VŠB – Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra kybernetiky a biomedicínského inženýrství

Interaktivní systém pro detekci řečového signálu

Interactive System with Speech Signal Detection

2015 Josef Kročil

VŠB - Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra kybernetiky a biomedicínského inženýrství

Zadání diplomové práce

Student: **Bc. Josef Kročil**

Studijní program: N2649 Elektrotechnika

Studijní obor: 3901T009 Biomedicínské inženýrství

Téma: Interaktivní systém pro detekci řečového signálu
Interactive System with Speech Signal Detection

Zásady pro vypracování:

1. Rozbor problematiky metod detekce řečového signálu.
2. Návrh metod provedení detekce řečového signálu.
3. Návrh a vývoj systému pro detekci řečového signálu.
4. Vizualizace a srovnání naměřených výsledků s teoretickými předpoklady.
5. Zhodnocení dosažených výsledků práce.

Seznam doporučené odborné literatury:

[1] BROUGHTON, S. Allen and Kurt M. BRYAN. *Discrete Fourier analysis and wavelets: applications to signal and image processing.* Hoboken, N.J.: Wiley, c2009, xv, 337 p. ISBN 978-0-470-29466-6.
[2] CLARK, Alexander, Chris FOX a Shalom LAPPIN. *The Handbook of Computational Linguistics and Natural Language Processing.* Malden, MA: Wiley-Blackwell, 2010, xxii, 775 p. ISBN 978-1-4051-5581-6.
[3] HUANG, Xuedong, Alex ACERO and Hsiao-Wuen HON. *Spoken language processing: a guide to theory, algoritm, and system development.* New Jersey: Prentice-Hall, 2001. 980 s. ISBN 0-13-022616-5.
[4] MCLOUGHLIN, Ian. *Applied Speech and Audio Processing.* Leiden: Cambridge University Press, 2009. 216 s. ISBN 978-0-521-51954-0.

Formální náležitosti a rozsah diplomové práce stanoví pokyny pro vypracování zveřejněné na webových stránkách fakulty.

Vedoucí diplomové práce: **Ing. Zdeněk Macháček, Ph.D.**

Datum zadání: 01.09.2013
Datum odevzdání: 07.05.2015

VYSOKÁ ŠKOLA BÁŇSKÁ-TECHNICKÁ UNIVERZITA Fakulta elektrotechniky a informatiky OSTRAVA

doc. Ing. Jiří Koziorek, Ph.D.
vedoucí katedry

prof. RNDr. Václav Snášel, CSc.
děkan fakulty

„Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal."

V Ostravě 7.5.2015 .....

Na tomto místě bych rád poděkoval Ing. Zdeňku Macháčkovi, Ph.D., za odborné vedení, rady a připomínky, které mi v průběhu psaní této diplomové práce poskytnul. Také děkuji lidem z mého okolí za možnost pořízení potřebných zvukových nahrávek.

Abstrakt

Tato diplomová práce se zabývá problematikou detekce řečového signálu a jeho zpracováním za účelem rozpoznání typu mluvčího (muž, žena, dítě), rozpoznání izolovaných slov a odhadu samohlásek *a*, *e*, *i*, *o*, *u*. Metody zpracování řečového signálu jsou provedeny v časové i frekvenční oblasti. Rozpoznání typu mluvčího je založeno na hledání základní periody hlasivek v kepstrální oblasti. Pro potřebu rozpoznání izolovaných slov (algoritmus DTW) jsou jako příznaky použity koeficienty LPC, které jsou rovněž dosazeny coby konstantní parametry do přenosové funkce hlasového traktu za účelem odhadu výše zmíněných samohlásek. Uživatelské interaktivní prostředí systému pro analýzu zvukového záznamu je vytvořeno ve výpočetním programu MATLAB. Výstupy jednotlivých analýz jsou znázorněny v grafické a také textové formě.

Klíčová slova

Řeč, detektor, kepstrum, lineární predikce, DTW, formanty.

Abstract

This thesis deals with the detection of speech and its processing in order to recognize the type of speaker (male, female, child), isolated word recognition, and an estimate of the vowels *a*, *e*, *i*, *o*, *u*. Methods of speech processing are carried out in time and frequency domain. Recognizing the type of speaker is based on finding fundamental period of vocal cords in cepstral area. To assist the recognition of isolated words (DTW algoritm), LPC coefficients are used, which are also used as constant parameters in the transfer function of the vocal tract in order to estimate the aforementioned vowels. Graphical user interface for the analysis of the audio recording is made in the computing program MATLAB. The outputs of each analysis are shown in graphical and text form.

Key Words

Speech, detector, cepstrum, linear prediction, DTW, formants.

Seznam použitých zkratek

ASR	automatické rozpoznání řeči (Automatic Speech Recognition)
DFT	diskrétní Fourierova transformace
DTW	dynamické borcení časové osy (Dynamic Time Warping)
HMM	skryté Markovovy modely (Hidden Markov Models)
IDFT	inverzní diskrétní Fourierova transformace
IPA	mezinárodní fonetická abeceda (International Phonetic Alphabet)
LAG	zpoždění
LPC	lineárně prediktivní kódování (Linear Predictive Coding)
MACF	modifikovaná autokorelační funkce (Modified Autocorrelation Function)
SAMPA	fonetická abeceda SAMPA (Speech Assesment Methods Alphabet)

Seznam použitých symbolů

$\boldsymbol{A}$	obraz testovaného slova
α	konstanta kepstrálního detektoru
$\boldsymbol{B}$	obraz referenčního slova
β	časová konstanta exponenciálního průměrování
$D(\boldsymbol{A},\boldsymbol{B})$	celková vzdálenost mezi obrazy $\boldsymbol{A}$ a $\boldsymbol{B}$
$d(n)$	aditivní šumové pozadí (diskrétní signál)
$d(\boldsymbol{x},\boldsymbol{v})$	zkreslení mezi vektory $\boldsymbol{x}$ a $\boldsymbol{v}$
Δc_l	kepstrální vzdálenost
E_d	úroveň energie hluku
E_n	krátkodobá energie signálu
E_p	prahová hodnota energie
$\varphi(\omega,n)$	fázové spektrum
F_0 (Hz)	základní frekvence hlasu
F_1 (Hz)	první formant
F_2 (Hz)	druhý formant
f_{vz} (Hz)	vzorkovací frekvence
f_m (Hz)	mezní (maximální) frekvence

$g(n,m)$	funkce lokálního omezení DTW
$g(t)$	impulsní periodický signál buzení (hlasivky)
G	koeficient zesílení
$G(f)$	amplitudová frekvenční charakteristika signálu buzení
$h(t)$	model impulsní odezvy řečového traktu
$H(f)$	amplitudová frekvenční charakteristika artikulačního traktu
$H_{WF}(f)$	přenosová charakteristika Wienerova filtru
$H(z)$	diskrétní model přenosové funkce řečového traktu
J	kriteriální funkce zkreslení
K	konstanta preemfáze
l	délka segmentu
M_k	krátkodobá intenzita signálu
N	celkový počet vzorků
$N(\hat{W})$	normalizační faktor DTW
p	parametr energetického nebo intenzitního detektoru
Q	řád lineární predikce
$R_n(m)$	krátkodobá autokorelační funkce
$s(n)$	diskrétní řečový signál
$s'(n)$	signál $s(n)$ po preemfázi
$s(t)$	spojitý řečový signál
$S(f)$	amplitudová frekvenční charakteristika hlasového ústrojí
$S_{vv}(f)$	výkonové spektrum šumu
$S_{xy}(f)$	vzájemné spektrum originálního a zkresleného signálu
$S_{yy}(f)$	výkonové spektrum zkresleného signálu
T_0 (s)	základní perioda hlasu
t (s)	čas
t_1	prahová úroveň kepstrálního detektoru
$y(n)$	řečový signál s aditivním šumem
$w(n)$	váhovací okno
Z_n	počet průchodů nulou

Obsah

1 Úvod

Zpracování řeči, tedy její počítačová analýza, nachází uplatnění v rámci řešení široké škály problémů souvisejících s komunikačními technologiemi. Jedná se zejména o detekci přítomnosti řeči v hlučném prostředí, o možnost převodu mluvené řeči na text, rozpoznání pohlaví osoby či přímo rozpoznání konkrétní osoby. V současné době jsou standardně k dispozici mnohé počítačové programy, které umožňují převést hlasový povel řečníka na definovaný úkon, např. stisknutí určité klávesy na klávesnici počítače apod. Zejména nevidomým osobám lze usnadnit např. tvorbu textových dokumentů, prostřednictvím převodu jejich řeči na text. Hlasové ovládání domácích spotřebičů není v současnosti úplným standardem, nicméně je stále vyvíjeno a zdokonalováno.

Náplň diplomové práce sestává z teoretické a praktické části. Druhá až šestá kapitola této diplomové práce je zaměřena na teoretický popis problematiky analýzy řeči. Jde především o vysvětlení mechanismu vzniku řeči a popis základních metody úpravy řečového signálu před jeho zpracováním. Dále jsou popsány metody analýzy v časové a frekvenční oblasti, typy detektorů řeči a klasifikační metody používané při rozpoznávání řeči. Závěrečné kapitoly teoretické části jsou zaměřeny na popis metod výpočtu základního tónu hlasivek a popis základních typů přepisu zvuků řeči do symbolické formy.

Sedmá kapitola shrnuje popis praktické realizace interaktivního systému pro detekci řeči, navrženého ve výpočetním programu MATLAB. Současně také shrnuje úspěšnost naměřených výsledků.

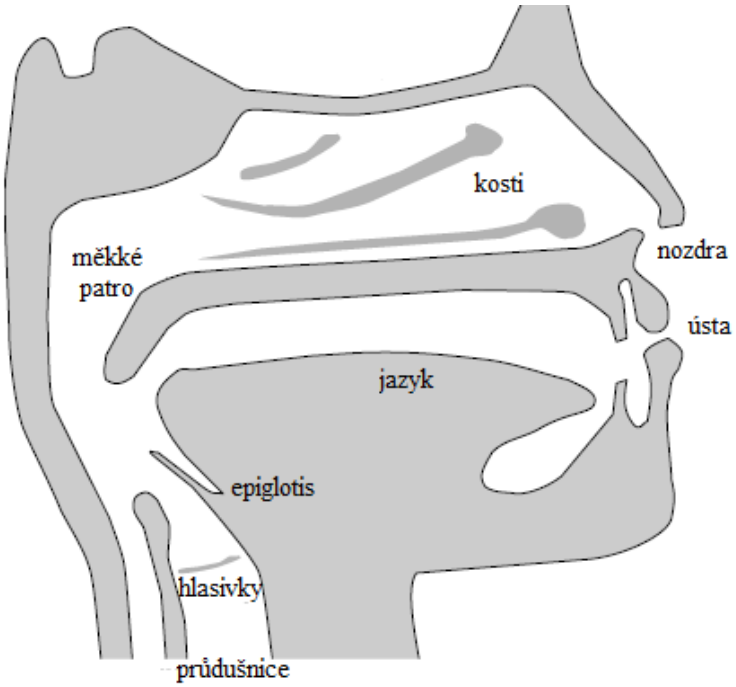
Cílem diplomové práce je na základě nastudovaných metod navrhnout systém pro detekci řeči ve formě zvukové nahrávky, s možností rozpoznání typu mluvčího (muž, žena) a rozpoznání vybraných hlásek či celých izolovaných slov. Naměřené výsledky musí být vyhodnoceny a porovnány s teoretickými předpoklady.

2 Řeč a její vznik

Lidská řeč je zřejmě nejdokonalejším dorozumívacím prostředkem z hlediska přenosu informace mezi inteligentními bytostmi. Řeč je charakterizována typickou akustickou strukturou, např. svým amplitudově-frekvenčním spektrem a subjektivním vlivem osobnosti řečníka, tedy barvou hlasu, intonací, apod. Nejmenší jednotkou řeči, podle které mohou být od sebe rozlišena jednotlivá slova, je foném, který lze zaznamenat např. u slov *šumět-čumět*. Tyto fonémy se po své zvukové stránce liší zejména v závislosti na místě jejich vzniku (některém z artikulačních orgánů), přičemž světové jazyky disponují přibližně 12 až 60 fonémů, v českém jazyce je jich 36. Zdrojem řečových kmitů, jakožto fyzikální podstaty vzniku řeči jsou řečové orgány, a sice hlasivky, dutina hrdelní, ústní, nosní, měkké a tvrdé patro, zuby a jazyk (obrázek 1). Znělé zvuky vznikají důsledkem kmitání hlasivek, respektive stažené hlasivkové štěrbiny, proudí-li přes ni vzduch z plic. Frekvence těchto kmitů je závislá na tlaku průchozího vzduchu a svalovém napětí hlasivek a je nazývána základní frekvencí F_0, která udává lidskému hlasu jeho základní tón, pomocí něhož lze klasifikovat mluvčího dle jeho pohlaví či věku. Hodnota základní frekvence se obecně pohybuje v rozmezí 50 až 400 Hz (tabulka 1).

Tabulka 1: Tabulka hodnot základního tónu hlasivek. [1]

Průměrné hodnoty základního tónu	
dětský hlas:	300-400 Hz
mužský hlas:	120 Hz
ženský hlas:	210 Hz



Obrázek 1: Artikulační ústrojí člověka. [2]

Při artikulaci samohlásek neboli vokálů je udržován co možná nejvolnější průchod vzduchu hlasivkami. V akustickém spektru samohlásek je možno pozorovat kromě základního hlasivkového tónu také řadu vyšších zesílených tónů, které jsou důsledkem rezonancí v dutinách hlasového traktu. Tyto tóny o vyšších frekvencích se nazývají formanty a bývají značeny písmenem *F*, s přiřazeným dolním indexem, reprezentujícím posloupnost velikosti frekvence jednotlivých formantů od nejnižší k největší, tedy $F_1, F_2,\ldots,F_n$. V rámci českého jazyka jsou za nejdůležitější považovány formanty F_1 a F_2 (tabulka 2), jejichž výška a intenzita vypovídá o aktuálním postavení hlasového traktu mluvčího, z hlediska geometrického.

Tabulka 2: Tabulka formantových frekvencí. [1]

Samohláska	**formant** F_1	**formant** F_2
u:	300-500 Hz	600-1000 Hz
o:	500-700 Hz	900-1200 Hz
a:	750-1100 Hz	1100-1500 Hz
e:	500-700 Hz	1500-2000 Hz
i:	300-500 Hz	2000-3000 Hz

Souhlásky neboli konsonanty se od prostých samohlásek liší zejména přítomností charakteristického šumu v akustickém spektru hlásek. Tyto konsonanty jsou produktem vzduchové turbulence, vznikající třením vydechovaného vzduchu o bariéru (překážku) tvořenou artikulačními orgány, např. jazykem, rty, apod. Tato bariéra může být buď úplná, nebo částečná. V případě úplné bariéry vzniká při šumových souhláskách v momentě zrušení této bariéry krátký, tzv. explozivní šum, podobající se svým charakterem náhlému výbuchu. Tyto jsou nazývány souhláskami závěrovými, tzv. okluzívami. V češtině jsou okluzívami: p, t, ť, k, b, d, ď, g, m, n, ň.

V případě bariéry částečné jde o zúžení cesty, jíž je vzduch vydechován, v některém místě artikulačního ústrojí. Tímto dochází ke vzniku třecího šumu a tvorbě souhlásek úžinových, tzv. frikativů (tabulka 3).

Tabulka 3: Tabulka českých frikativů. [1]

Frikativy	
vlastní úžinové	české f, v, s, š, z, ž, j, ch, h
bokové (lateráry)	české l
kmitavé (vibranty)	české r, ř

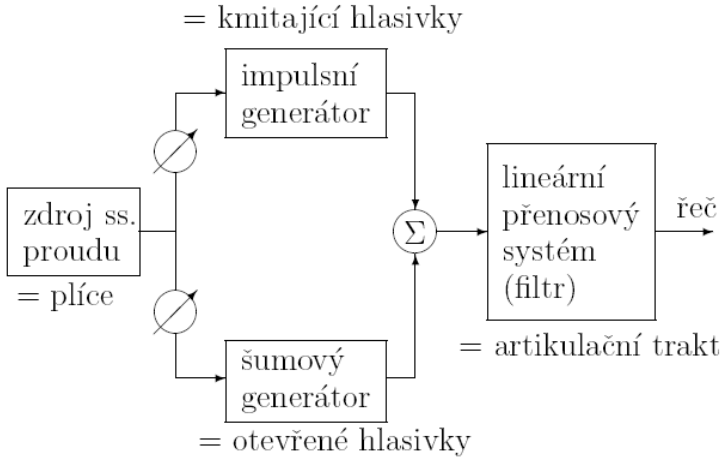
Malá skupina hlásek disponuje oběma typy bariér. Jedná se o souhlásky polozávěrové, tzv. semiokluzívy a zahrnují české c a č.

Důležitým hlediskem pro klasifikaci souhlásek je jejich znělost. Při vyslovování neznělých souhlásek je stav hlasivek proporcionálně podobný stavu při volnému výdechu, dochází k volnému propouštění vydechovaného vzduchu, bez dodatečné tvorby hlasu. Znělé souhlásky, jak již bylo řečeno, disponují přítomností základního tónu, respektive frekvence F_0. Nosní dutina hraje roli při artikulaci nosních souhlásek m, n, ň. Šumové souhlásky lze seřadit do dvojic, které se neliší

způsobem artikulace, liší se však účastí hlasu, čili znělostí. Takové souhlásky jsou nazývány párové. Nepárové souhlásky jsou vždy znělé a nedisponují neznělým protějškem.

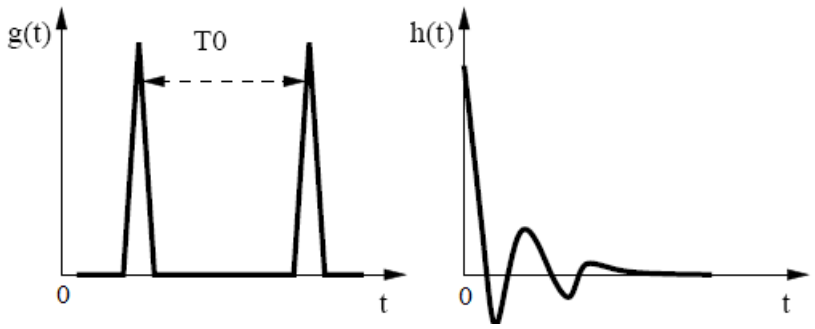
[1],[2]

Na základě předešlých informací lze definovat model, resp. blokové schéma mechanismu vzniku řeči (obrázek 2). Pomocí elektronické analogie (řečový syntetizér) lze definovat plíce coby zdroj proudu vzduchu jako zdroj stejnosměrného el. proudu, kmitající hlasivky (přítomnost F_0) lze považovat za ekvivalent impulsního generátoru a otevřené hlasivky (neznělé) mohou zde být považovány za šumový generátor. Artikulační trakt v tomto případě hraje roli lineárního filtru.



Obrázek 2: Blokové schéma řečového syntetizéru. [4]

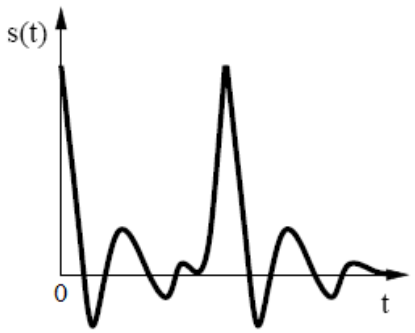
V časové oblasti lze reprezentovat funkci hlasivek impulsním periodickým signálem $g(t)$, coby buzením, o základní periodě T_0, vlastnosti artikulačního traktu pak signálem $h(t)$, jenž je impulsní odezvou výše zmíněného filtru (obrázek 3). Tvorbu řeči lze v tomto případě popsat jako průchod signálu $g(t)$ přes filtr s impulsní odezvou $h(t)$.



Obrázek 3: Model signálu buzení (vlevo) a časový model impulsní odezvy artikulačního traktu (vpravo). [4]

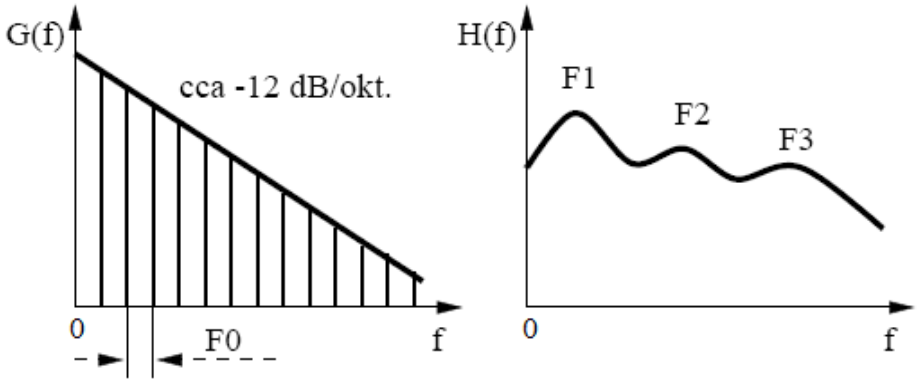
Výsledný model řečového signálu $s(t)$ (obrázek 4) je tedy dán konvolucí buzení a dané impulsní odezvy v čase.

$$s(t) = g(t) \star h(t) = \int_{-\infty}^{\infty} g(t)h(t-\tau)d\tau \qquad (1)$$



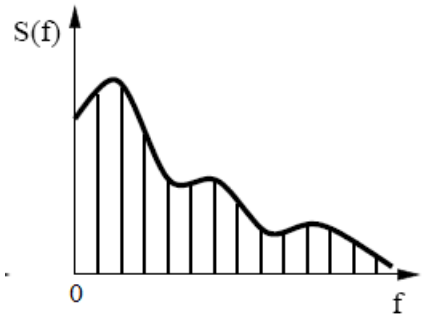
Obrázek 4: Model řečového signálu znělé hlásky. [4]

Ve frekvenční oblasti má periodický signál $g(t)$ čárové amplitudové spektrum, impulsní charakteristika $h(t)$ disponuje po Fourierově transformaci spektrálními vlastnostmi filtru, s viditelnými lokálními maximy, vztahujícími se k frekvencím jednotlivých formantů (obrázek 5).



Obrázek 5: Frekvenční charakteristika signálu buzení (vlevo) a artikulačního traktu (vpravo). [4]

Vynásobením amplitudového spektra $G(f)$ a $H(f)$ lze získat výslednou frekvenční charakteristiku $S(f)$ řečového signálu (obrázek 6).



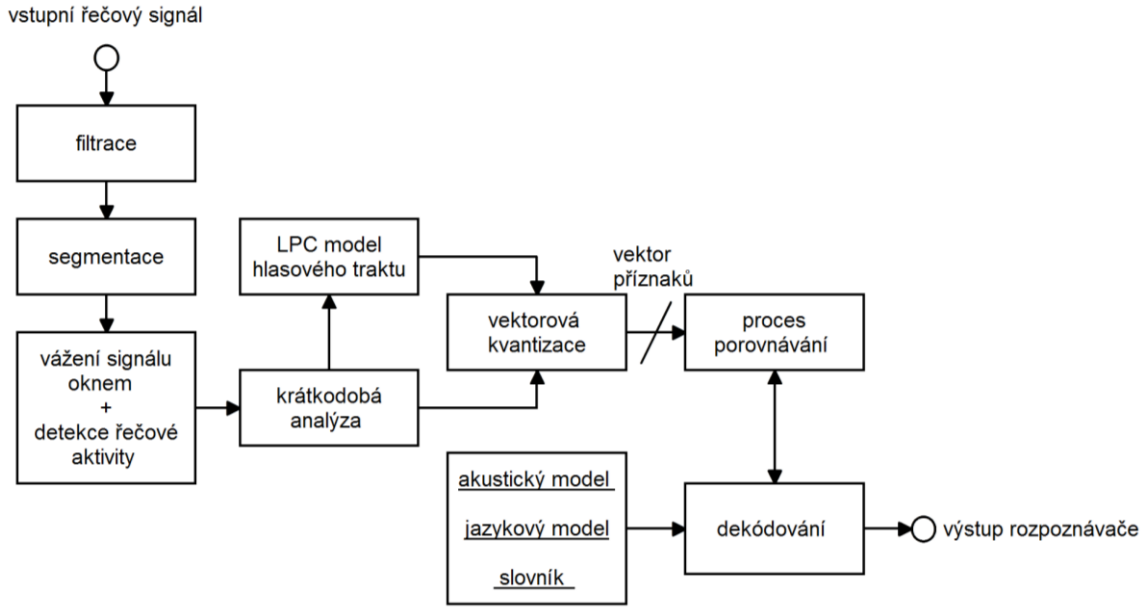
Obrázek 6: Model frekvenční charakteristiky hlasového ústrojí. [4]

Pro oddělení složek buzení a artikulačního traktu (např. pro zjištění základního tónu hlasivek) je nutno provést zpětnou dekonvoluci. Jednou z metod nepřímé dekonvoluce je kepstrální analýza, jenž je účinným, nicméně výpočetně náročným matematickým nástrojem odvozeným z diskrétní Fourierovy transformace a ve zpracování řečových signálů bývá velmi často používán. Tato metoda bude později popsána a aplikována v rámci této diplomové práce.

[4]

3 Základní metody zpracování řečového signálu

Přístupy ke zpracování řečového signálu pro potřebu detekce a rozpoznání řeči mohou být dvojího typu. První možností je analýza dat, která jsou k dispozici ve formě hotového záznamu (tzv. offline přístup), druhou možností je kontinuální pořizování zvukové nahrávky, přičemž výsledky analýzy takovéto nahrávky jsou zpracovávány, ukládány a aktualizovány v krátkých časových intervalech (tzv. online přístup). Pro oba zmiňované přístupy lze zavést jako příklad principielní zjednodušené blokové schéma řetězce pro automatickou detekci a rozpoznání řečového signálu (zkr. ASR, z angl. automatic speech recognition).



Obrázek 7: Blokové schéma systému ASR (převzato a upraveno). [2]

Blokové schéma systému ASR (obrázek 7) znázorňuje posloupnost úkonů reprezentujících proces rozpoznání řeči. Vstupní signál je nejprve filtrován a v krátkých časových úsecích (segmentech) následně vyhodnocován detektorem řečové aktivity. Je li v těchto segmentech přítomna řeč, pak pro každý takový segment následují výpočty krátkodobých charakteristik v čase a frekvenci, často včetně tzv. LPC odhadu modelu spektra hlasového traktu, který popisuje aktuální pozici hlasového traktu při promluvě řečníka. Z těchto údajů je pro daný segment sestaven vektor příznaků, který je porovnáván se vzorovými modely. Na základě největší shody s některým vzorem dané třídy je pak příchozí vektor přiřazen do téže třídy. Poté je dekódován např. na příslušný znak abecedy z databáze slovníku. Výstupem může být řetězec po sobě jdoucích fonémů, nebo v lepším případě sekvence slov. V následujících kapitolách teoretické části této diplomové práce jsou výše jmenované metody popsány podrobněji. [2]

3.1 Filtrace řečového signálu

Před samotným zpracováním řečového signálu je nejprve potřeba provést jeho akvizici. Podstata snímání řečového signálu tkví v převedení jisté, v čase proměnné fyzikální veličiny, např. akustického tlaku, na veličinu elektrickou, tedy elektrické napětí. Toho je dosaženo pomocí mikrofonu. Poté je zapotřebí tento elektrický signál zesílit, neboť napětí, generované mikrofonem dosahuje úroveň cca několika milivoltů. Aby mohl být měřený signál analyzován, je nutno jej převést z analogové do digitální podoby (vzorkování a kvantizace) pomocí A/D převodníku a za pomoci dalších metod následně v určitém datovém formátu uložit do paměťového média či přímo počítače ke zpracování. S ohledem na Shannonův vzorkovací teorém bývá v měřicím řetězci zapojen anti-aliasingový filtr typu dolní propust, s mezní frekvencí f_m menší, než jedna polovina vzorkovací frekvence f_{vz}.

$$f_{vz} > 2f_m \tag{2}$$

Pro potlačení síťového rušení a vlastního rušení mikrofonu může být při nahrávání použit filtr typu horní propust s mezním kmitočtem asi do 200 Hz, a to buď jako součást analogového měřicího řetězce, či v podobě digitální. Další z možností je použití filtru pásmová propust nebo kombinací tzv. banky těchto filtrů. Pro potlačení stacionárního šumu bývají v praxi aplikovány jednodušší metody, založené na spektrálním odečítání. Je třeba poznamenat, že filtrace řečového signálu v jeho užitečném pásmu s sebou někdy může nést rizika ztráty některých, v původním signálu obsažených informací.

[1], [2], [5]

3.1.1 Spektrální odečítání

Základním způsobem odstranění vlivu rušivého prostředí, majícího vlastnost stacionárního šumu je spektrální odečítání. Předpokladem pro použití metody je nutnost použití řečového detektoru, který během řečových pauz aktualizuje odhad spektrální výkonové hustoty šumu. Smíšení řečového signálu se šumem lze popsat vztahem

$$y(n) = s(n) + d(n)\,, \tag{3}$$

kde $s(n)$ je původní řečový signál a $d(n)$ je aditivní šum pozadí.

Spektrální výkonová hustota obnoveného řečového signálu $\left|\hat{S}(\omega)\right|^2$ je pak dána odečtením odhadu šumu $\left|\hat{D}(\omega)\right|^2$ od odhadu spektrální výkonové hustoty degradované řeči $|Y(\omega)|^2$ při dodržení podmínek daných dle vztahu

$$\left|\hat{S}(\omega)\right|^2 = \begin{cases} |Y(\omega)|^2 - \left|\hat{D}(\omega)\right|^2 & ,pro\ |Y(\omega)|^2 > \left|\hat{D}(\omega)\right|^2 \\ 0 & ,jinak \end{cases}. \tag{4}$$

Posledním krokem vedoucím k obnovení řečového signálu jeho převedení do časové oblasti dle vztahu

$$\hat{s}(n) = Re\{IFFT(|\hat{S}(\omega)| \cdot e^{jarg[Y(\omega)]})\}, \quad (5)$$

kde $arg[Y(\omega)]$ je fáze původního zašuměného signálu.

Existuje několik modifikací, které mohou zlepšit účinnost spektrálního odečtu a zamezit vzniku hudebních tónu v obnoveném signálu. Jsou jimi zejména Beroutiho algoritmus a algoritmy založené na doplnění o tzv. ziskovou *G* funkci.

[5], [6]

3.1.2 Wienerova filtrace

Wienerův filtr, stejně jako metoda spektrálního odečítání, může být použit pro odstranění stacionárního šumu. Přenosovou charakteristiku Wienerova filtru lze odvodit z požadavku minimální kvadratické chyby mezi řečovým signálem a odhadem téhož signálu. Přenosová charakteristika je dána vztahem

$$H_{WF}(f) = \frac{S_{xy}(f)}{S_{yy}(f)}, \quad (6)$$

kde $S_{xy}(f)$ je vzájemné spektrum originálního a zkresleného signálu a $S_{yy}(f)$ je výkonové spektrum zkresleného signálu. Jsou li řečový signál a rušení fázově ortogonální, což v mnoha případech platí, pak

$$S_{xy}(f) = S_{xx}(f), \quad (7)$$

$$S_{yy}(f) = S_{xx}(f) + S_{vv}(f), \quad (8)$$

kde S_{xx} je výkonové spektrum originálního signálu a S_{vv} výkonové spektrum šumu. Rovnici (6) lze poté přepsat do tvaru

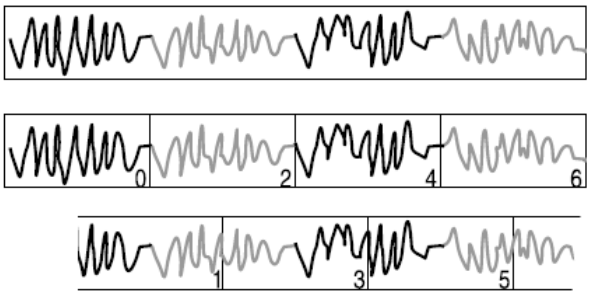
$$H_{WF}(f) = \frac{1}{1 + \frac{S_{vv}(f)}{S_{xx}(f)}}. \quad (9)$$

Předešlý vztah popisuje tzv. nekauzální Wienerův filtr, poněvadž v praxi obvykle informace o výkonovém spektru originálního signálu není k dispozici. Tento problém může být řešen Kalmanovou filtrací.

[7]

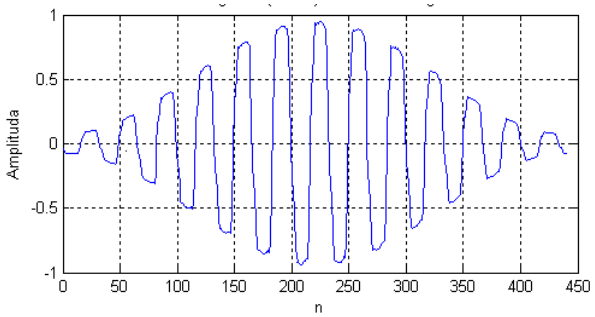
3.2 Segmentace, preemfáze, normalizace

Vzhledem k tomu, že řečový signál jako celek je brán jako signál stochastický, musí být rozdělen do tzv. segmentů neboli rámců takové délky, aby se parametry signálu v tomto úseku měnily velmi pomalu. S ohledem na rychlost průběhu změn parametrů artikulačního ústrojí jsou délky segmentů voleny nejběžněji v rozmezí 15-30 ms. Všechny segmenty přitom disponují stejnou délkou. Segmentace může být udělána formou prostého dělení nahrávky na rámce délky *l* vzorků z celkového počtu *N* vzorků, popřípadě se mohou segmenty vzájemně překrývat (obrázek 8).



Obrázek 8: Řečový signál (nahoře), jeho segmentace bez překrytí (uprostřed) a s poloviční překrytím (dole). [2]

K potřebné extrakci segmentů z nahrávky se používají tzv. okénkové funkce, které jsou na stanoveném intervalu (v tomto případě šířce segmentu) reprezentovány buď konstantní hodnotou (obdélníkové okno), nebo nějakou funkcí (Hammingovo okno), nabývající na tomto intervalu určitých funkčních hodnot. Mimo stanovený interval nabývají okénkové funkce nulových hodnot. Vzájemným násobením posunující se okenní funkce a vektoru hodnot amplitud vzorků jsou vybírány rámce a je jim udělována tzv. váha, závisející na parametrech této okenní funkce. Není-li celkový získaný počet rámců vyjádřen celým číslem, pak poslední rámec, který nedisponuje potřebným počtem vzorků je jednoduše vyřazen z dalšího procesu zpracování. Většinou se používá Hammingovo okno (obrázek 9), které je sice méně selektivní, než okno obdélníkové, avšak neznehodnotí spektrum segmentu (rozptyl spektra a tvorba poměrně velkých postranních laloků na vyšších frekvencích), jak je tomu u okna obdélníkového (obrázek 10). [2], [4], [1]

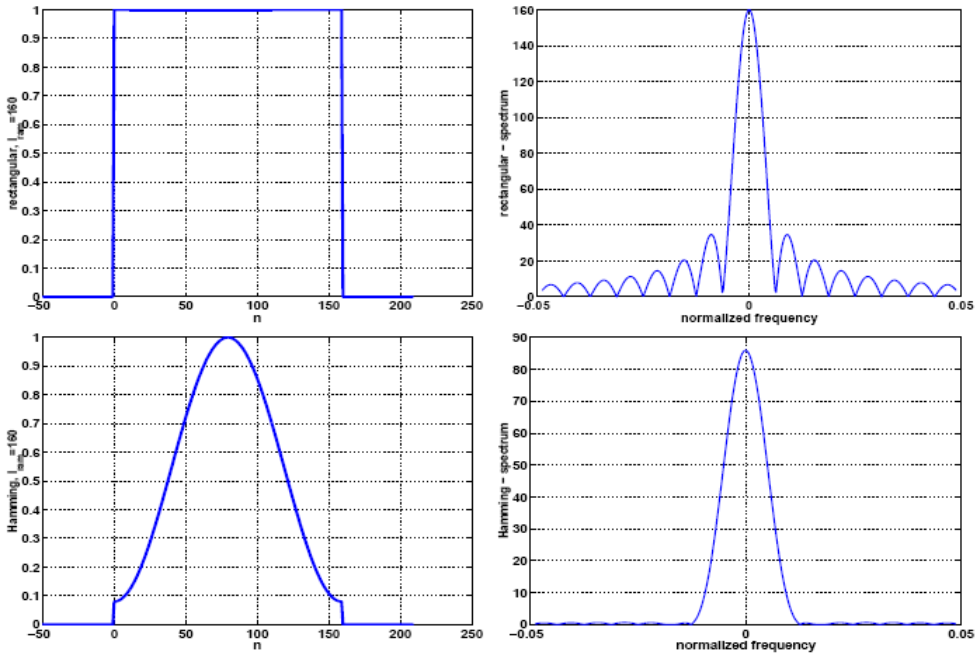


Obrázek 9: Segment sinusového signálu vážený Hammingovým oknem.

Hammingovo okno $w(n)$ je definováno následujícím matematickým vztahem

$$w(n) = 0{,}54 - 0{,}46\cos\left(2\pi\frac{n}{N}\right) \text{ pro } 0 \leq n \leq N-1\,, \qquad (10)$$

kde n je n-tý vzorek váženého diskrétního signálu a N je celkový počet vzorků váženého diskrétního signálu.



Obrázek 10: Obdélníkové okno v časové (vlevo nahoře) a frekvenční oblasti (vpravo nahoře) a tytéž charakteristiky pro okno Hammingovo (vlevo a vpravo dole). [4]

Před samotným vážením signálu je možno provést tři užitečné operace, jimiž jsou ustředění, normalizace a preemfáze signálu. Řečový signál lze ustředit prostým odečtením jeho střední hodnoty od každého vzorku. Tím je signálu odebrána stejnosměrná složka. Následná normalizace udělí signálu maximální rozsah amplitudy od -1 do 1. Ta je výsledkem prostého podílu každého vzorku signálu s maximem z absolutní hodnoty tohoto signálu. V rámci zvýraznění vyšších frekvencí (formantů, ale často bohužel i šumu) lze aplikovat tzv. preemfázi, která je dána následující diferenční rovnicí

$$s'(n) = s(n) - K \cdot s(n-1)\,, \qquad (11)$$

kde $s'(n)$ je hodnota n-tého vzorku signálu $s(n)$ po preemfázi a K je konstanta taková, jejíž rozsah je v mezích $0 \leq K \leq 1$. Ve své podstatě se jedná o číslicový filtr prvního řádu typu horní propust.

3.3 Krátkodobá analýza v časové oblasti

Mezi základní přístupy k analýze řečových signálů v čase patří zejména výpočet krátkodobé energie, popř. intenzity. Tyto veličiny slouží hlavně k detekci řečového signálu v nahrávce a lze je použít i k oddělení znělých a neznělých částí promluvy. Další často používané metody jsou autokorelace a výpočet počtu průchodů signálu nulou.

3.3.1 Střední počet průchodů signálu nulou

Často používanou veličinou je krátkodobá funkce počtu průchodů signálu nulou, daná vztahem

$$Z_n = \sum_{k=-\infty}^{\infty} |sgn(s[k]) - sgn(s[k-1])|\, w(n-k)\,, \tag{12}$$

kde *sgn* je znaménková funkce

$$sgn[s(k)] = \begin{Bmatrix} 1\, pro\, s(k) \geq 0 \\ -1\, pro\, s(k) < 0 \end{Bmatrix}, \tag{13}$$

s(*k*) je hodnota *k*-tého vzorku veličiny *s* a *w*(*n*) je váhovací okno.

Na základě počtu průchodů signálu nulou lze klasifikovat znělost hlásek a při určité modifikaci této funkce ji lze také použít pro určení základního tónu hlasivek či přibližnému odhadu frekvencí formantů. Jde tedy o jednoduchou charakteristiku, schopnou částečně popsat spektrální vlastnosti signálu.

3.3.2 Autokorelační funkce

Krátkodobá autokorelační funkce signálu o délce N vzorků je definována následujícím vztahem

$$R_n(m) = \sum_{k=-\infty}^{\infty} s(k)w(n-k)s(k+m)w(n-k-m)\,, \tag{14}$$

kde *s*(*k*) je hodnota *k*-tého vzorku korelovaného signálu, *m* je hodnota posunu téhož signálu o *m* vzorků a *w*(*n*) je váhovací okenní funkce.

Pokud je zpracovávaný signál periodický s periodou *P*, pak autokorelační funkce nabývá maximálních hodnot pro *k* = 0, *P*, 2*P*, atd. Pomocí této jednoduché analýzy lze určit např. periodu základního tónu. Podmínkou je ovšem dostatečná délka řečového segmentu, neboť musí obsahovat alespoň dvě periody signálu.

3.3.3 Energie signálu

Energii signálu lze vypočítat dle vztahu

$$E_n = \sum_{k=-\infty}^{\infty} [s(k)w(n-k)]^2\,, \tag{15}$$

kde *s*(*k*) je *k*-tý vzorek signálu *s* a *w*(*n*) je některým z mnoha typů váhovacích oken (většinou Hamming).

[1], [4]

3.4 Krátkodobá analýza ve frekvenční oblasti

Krátkodobá Fourierova transformace *S*(*ω*, *n*) je definována vztahem

$$S(\omega,n)=\sum_{k=-\infty}^{\infty}s(k)w(n-k)e^{-j\omega k}\,, \tag{16}$$

kde $w(n)$ je příslušný typ váhovacího okénka, $s(k)$ je vzorek diskrétního signálu.

Výsledkem transformace je obecná komplexní funkce ve tvaru

$$S(\omega,n)=a(\omega,n)+jb(\omega,n)\,, \tag{17}$$

kde $a(\omega,n)$ je reálná část komplexního čísla, $b(\omega,n)$ je imaginární část komplexního čísla.

V případě potřeby získání amplitudového spektra může být získán modul komplexní funkce dle vztahu

$$|S(\omega,n)|=\sqrt{a^2(\omega,n)+b^2(\omega,n)}\,. \tag{18}$$

Fázové spektrum je definováno následovně:

$$\varphi(\omega,n)=\operatorname{arctg}\frac{Im\{S(\omega,n)\}}{Re\{S(\omega,n)\}}, \tag{19}$$

kde $\varphi(\omega,n)$ je fázový úhel.

[1]

3.5 Lineární prediktivní analýza

Lineární prediktivní analýza, nebo také lineárně prediktivní kódování (zkr. *LPC*) je metoda, založená na principu odhadu modelu tvorby řeči z krátkodobých segmentů řečového signálu. Výhodou této metody je také přijatelná výpočetní zátěž a poměrně přesné odhady hledaných parametrů. Předpokladem použití lineární predikce je vyjádření k-tého vzorku řečového signálu $s(k)$ jako lineární kombinace Q předchozích výstupních vzorků a buzení $u(k)$, tj. vlivu hlasivek jakožto impulsního generátoru. Matematické vyjádření je popsáno následujícím vztahem

$$s(k)=-\sum_{i=1}^{Q}a_i s(k-i)+Gu(k)\,. \tag{20}$$

G je koeficient zesílení a Q je řád modelu, tedy řád prediktoru. Použitím Z transformace lze pak odvodit přenosovou funkci modelu $H(z)$, která je dána vztahem

$$H(z)=\frac{G}{1+\sum_{i=1}^{Q}a_i z^{-i}}, \tag{21}$$

kde sledovanými parametry jsou koeficienty a_i číslicového filtru a zesílení G. V případě dodržení stacionarity řečového signálu (tedy jeho segmentací na krátkodobé úseky) lze pro určení koeficientů filtru a jeho zesílení použít metodu nejmenších čtverců. Je li člen $Gu(k)$ v rovnici (20) neznámý, je z rovnice vypuštěn a vzniká tzv. chyba predikce $e(k)$, reprezentovaná sumou čtverců

z rozdílu mezi skutečnou hodnotou vzorku $s(k)$ a předpovězenou $\hat{s}(k)$. Ve své podstatě se jedná o krátkodobou energii chyby signálu dle vztahu

$$E = \sum_k e^2(k) = \sum_k [s(k) - \hat{s}(k)]^2 = \sum_k \left[s(k) + \sum_{i=1}^{Q} a_i s(k-i) \right]^2 . \tag{22}$$

Parciálními derivacemi kriteriální funkce podle všech a_i lze chybu minimalizovat (položit tyto derivace do rovnosti k nule). Úpravami těchto rovnic lze dospět k maticovému zápisu

$$\begin{bmatrix} R_n(0); & R_n(1); & R_n(2); & \dots; & R_n(Q-1); \\ R_n(1); & R_n(0); & R_n(1); & \dots; & R_n(Q-2); \\ \vdots & & & & \\ R_n(Q-1); & R_n(Q-2); & R_n(Q-3); & \dots; & R_n(0); \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_Q \end{bmatrix} = \begin{bmatrix} -R_n(1) \\ -R_n(2) \\ \vdots \\ -R_n(Q) \end{bmatrix} \tag{23}$$

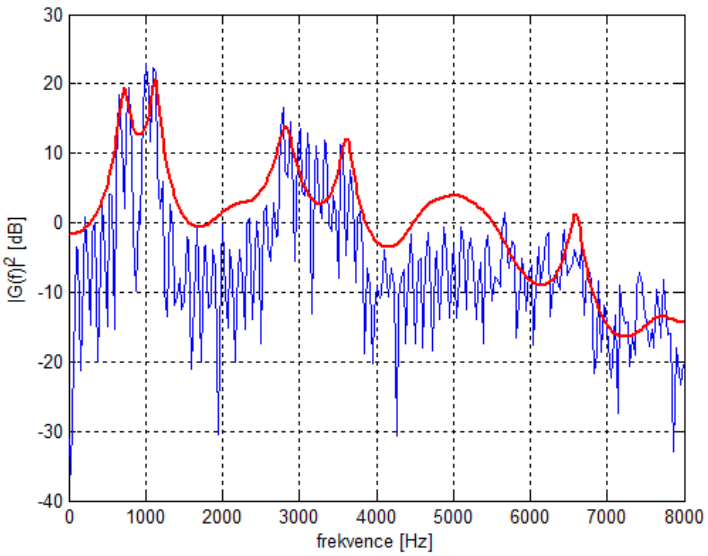
kde $R_n(0)$ až $R_n(Q)$ jsou autokorelační koeficienty n-tého segmentu řečového signálu. Matice na levé straně je symetrická, v tzv. Töplitzově tvaru. Minimální střední kvadrát chyby predikce v krátkodobém intervalu lze vyjádřit vztahem

$$E_n = R_n(0) + \sum_{i=1}^{Q} a_i R_n(i) . \tag{24}$$

Je li budicí signál ve tvaru jednotkového impulsu (znělá řeč) či bílého šumu (nezněná řeč), pak lze zesílení G popsat vztahem

$$G = \sqrt{E_n} . \tag{25}$$

Celý výpočet koeficientů predikce a zesílení lze významně zefektivnit iterativní metodou dle Levinsona a Durbina. Metodou lineární predikce je možno modelovat odhad signálového spektra řeči (obrázek 11), popřípadě lze z výsledků predikce řeč syntetizovat. [1]

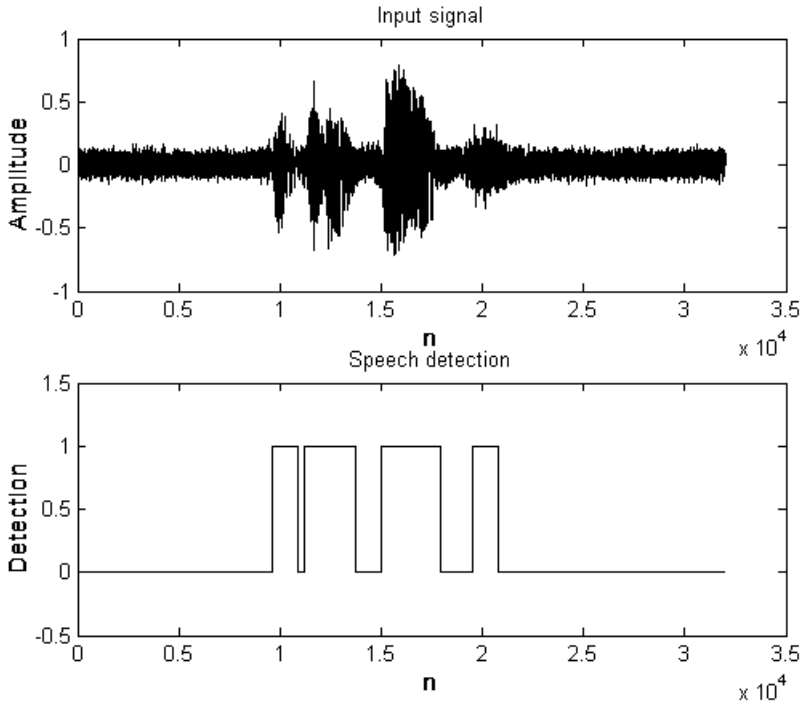


Obrázek 11: *LPC* odhad výkonového spektra segmentu hlásky „a“ (červeně).

4 Detektory řeč/pauza

4.1 Standardní detektory

Aby mohl být řečový signál podroben analýze, je třeba nejprve rozlišit hranice mezi promluvou a šumem pozadí. Detektor řeči je systém rozhodující o přítomnosti, popř. nepřítomnosti řeči v daném signálu. Výstupem takového detektoru je logická hodnota 0 (řeč nedetekována) nebo 1 (řeč detekována), která je přiřazena danému segmentu (obrázek 12). Segmenty, kde byla potvrzena přítomnost řeči, jsou poté ukládány k dalšímu zpracování. Základním požadavkem na řečový detektor je co možná největší účinnost i při malých poměrech signál/šum. [5]



Obrázek 12: Řečový signál (nahoře) a jeho detekce (dole). [5]

4.1.1 Energetické a intenzitní detektory

Energetické detektory jsou založeny na výpočtu krátkodobé energie jednotlivých segmentů dle vztahu

$$E_k = \sum_{k=-\infty}^{\infty} [s(k)w(n-k)]^2 , \tag{26}$$

kde $s(k)$ je k - tý vzorek signálu s a $w(n)$ je příslušný typ váhovacího okna.

Alternativou k výpočtu energie je výpočet intenzity signálu, která nemá tak velký dynamický rozsah hodnot (důsledek umocnění na druhou) a je méně citlivá na velké změny úrovně signálu. Krátkodobá intenzita signálu je dána vztahem

$$M_k = \sum_{k=-\infty}^{\infty} |s(k)|w(n-k)\,. \tag{27}$$

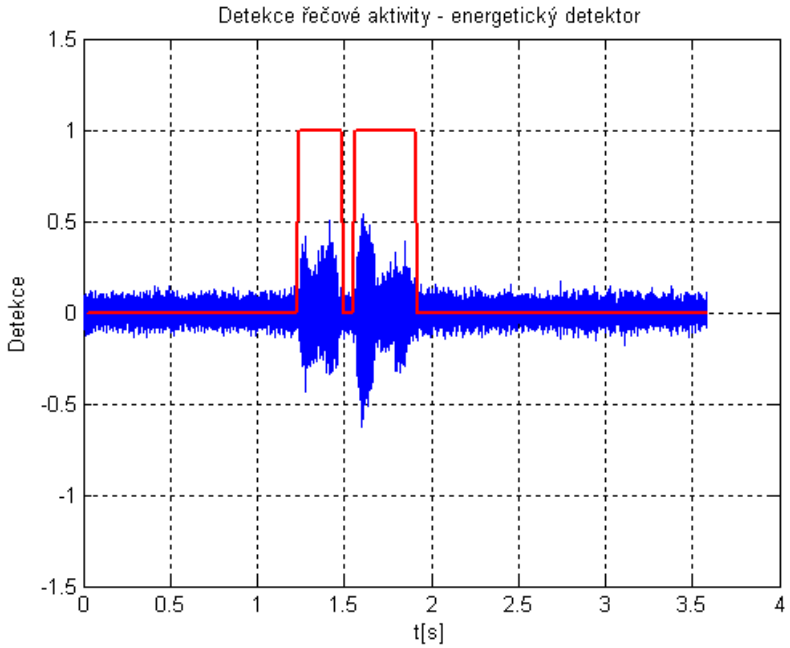
Algoritmus funkce energetického detektoru je založen na porovnání hodnoty krátkodobé energie aktuálního segmentu vypočteného dle rovnice (26) s prahovou hodnotou energie E_p, vypočtenou dle vztahu

$$E_p = 1{,}5E_d\,, \tag{28}$$

kde E_d je úroveň hluku pozadí, aktualizována podle vztahu

$$E_d^{aktuální} = (1-p)E_d^{předchozí} + pE_k\,, \tag{29}$$

kde p je číselný parametr, obvykle roven hodnotě 0,5. Aktualizace předchozího vztahu se provádí jen v řečových pauzách. Pokud je hodnota energie E_k větší než E_p, pak je výstup detektoru roven jedné. V opačném případě nerovnosti je výstupem detektoru hodnota 0 a energie hluku pozadí je aktualizována dle vztahu (29). Na obrázku (obrázek 13) je znázorněn průběh detekce promluvy věty „ahoj Karle“ v hlučném prostředí (fén). [5], [1]



Obrázek 13: Detekce řečového signálu energetickým detektorem.

4.1.2 Kepstrální integrální detektor

Kepstrální detektor využívá pro svou funkci reálnou část komplexního kepstra. Pro l - tý segment z N vzorků signálu lze reálné kepstrum popsat vztahem

$$c_l[k] = Re\{IDFT\{log|DFT\{s_l(n)\}|\}\} \tag{30}$$

kde $s_l(n)$ je řečový signál l-tého segmentu.

Detekce řeči u kepstrálního detektoru je založena na výpočtu tzv. kepstrální vzdálenosti Δc_l, dle vztahu

$$\Delta c_l = 4{,}3429 \sqrt{\left((c_1[0] - \overline{c_l}[0])^2 + 2 \sum_{k=1}^{p} (c_1[k] - \overline{c_l}[k])^2 \right)}, \tag{31}$$

kde $\overline{c_l}[k]$ je střední hodnota kepstra pozadí. Pokud platí nerovnost $\Delta c_l \geq t_1$, bude na výstupu detektoru přítomna hodnota 1. V opačném případě nerovnosti nebude detekována řeč a výstup detektoru bude disponovat hodnotou 0. Parametr t_1 je prahová úroveň, aktualizovaná v řečových pauzách, stejně jako v případě energetického detektoru. Tato prahová úroveň je určena statisticky, dle vztahu

$$t_1 = E(\Delta c[n]) + \propto \sqrt{D(\Delta c[n])}\,, \tag{32}$$

kde $E(\Delta c[n])$ je střední hodnota kepstrální vzdálenosti, α je volitelná konstanta o hodnotě 1,5 až 3,5 a výraz $\sqrt{D(\Delta c[n])}$ je hodnota směrodatné odchylky kepstrální vzdálenosti.

Aktualizace průměrného kepstra šumového pozadí je prováděna prostřednictvím vztahu

$$\overline{c_{l+1}}[k] = (1 - \beta)\overline{c_l}[k] + \beta c_l[k]\,, \tag{33}$$

kde β je časová konstanta exponenciálního průměrování, v rozmezí od 0 do 1.

Tento typ detektoru potřebuje pro správnou činnost tzv. počáteční prodlevu neboli inicializační fázi, během které je z několika počátečních segmentů spočtena střední hodnota a rozptyl kepstrální vzdálenosti. [5]

4.2 *HMM* detektory

Dalším typem řečových detektorů jsou detektory pracující na principu tzv. skrytých Markovových modelů (zkr. HMM, angl. Hidden Markov Model). Tyto jsou založeny na představě vytváření řeči člověkem a jejich snahou je pomocí statistické pravděpodobnosti určit míru pravděpodobnosti šumu nebo řečového signálu. Tato diplomová práce se zabývá pouze detektory standardními, proto již nebudou detektory využívající Markovovy modely dále rozebírány. [1]

5 Klasifikace a rozpoznání příznaků řeč. signálu

5.1 Vektorová kvantizace

Kvantizace je proces, jehož cílem je aproximovat analogovou hodnotu nějaké veličiny právě jednou hodnotou z konečného počtu číselných hodnot. Je – li prováděna kvantizace jednotlivé signálové veličiny či parametru, jde o kvantizaci skalární. Jestliže je kvantizován spojený blok veličin či paramentrů jako celek, pak jde o kvantizaci blokovou nebo také vektorovou. Této se využívá zejména při zpracování dat (příznaků) extrahovaných z jednotlivých segmentů řečového signálu. Vektorová kvantizace může být použita pro kompresi dat nebo v oblasti klasifikačních procesů. Matematická formulace vektorové kvantizace je následující: nechť existuje Q dimenzionální prostor X, v němž je dána množina vektorů $\boldsymbol{x}=[x_1,x_2,\ldots,x_Q]^T$, přičemž $\boldsymbol{x} \in X$. Komponenty vektorů x_i $(1 \leq i \leq Q)$ jsou reálné hodnoty náhodně proměnných se spojitou amplitudou. Q - dimenzionální kvantizér o L úrovních pak přiděluje každému vstupnímu vektoru $\boldsymbol{x}$ tzv. reprodukční vektor $\boldsymbol{v} = q(\boldsymbol{x})$, jenž je vybrán z konečné reprodukční abecedy $\boldsymbol{V} = \{v_1,\ldots,v_L\}$, přičemž vektory $\boldsymbol{v}_i$ mají diskrétní amplitudu a jsou ve většině případů rovněž Q - dimenzionální**.** Kvantizér q je tedy popsán reprodukční abecedou neboli kódovou knihou $\boldsymbol{V}$ o L úrovních a dělí prostor X na L disjunktních oblastí X_i. Každá oblast X_i je spojena s vektorem $\boldsymbol{v_i}$ dle vztahu

$$q(\boldsymbol{x}) = \boldsymbol{v}_i, \boldsymbol{x} \in X_i \,. \tag{34}$$

Kvantizací vstupního vektoru $\boldsymbol{x}$, tedy jeho nahrazením kódovým vektorem $\boldsymbol{v}$, vzniká kvantizační chyba nebo také zkreslení. Toto zkreslení lze kvantifikovat měrami vzdálenosti či zkreslení $d(\boldsymbol{x},\boldsymbol{v})$, kde druh míry je volen v závislosti na typu kvantizovaných veličin. Každý kódový vektor $\boldsymbol{v_i}$ musí být určen tak, aby průměrné zkreslení v oblasti X_i bylo minimální. Takový vektor $\boldsymbol{v}_i$ je pak tzv. centroidem oblasti X_i. V praxi většinou není známa apriorní pravděpodobnost, ani funkce hustoty pravděpodobnosti výskytu vektoru $\boldsymbol{x}$ v oblasti X_i. Je – li k dispozici konečný počet N vektorů $\boldsymbol{x}$ z tzv. trénovací množiny T a je – li znám cílový počet L položek kódové knihy, lze provést rozklad trénovací množiny T na podmnožiny, tzv. shluky T_i $(i = 1,\ldots, L)$ tak, aby byla minimalizována kriteriální funkce celkového zkreslení J dle vztahu

$$J = \min_{\boldsymbol{v}_i} \sum_{i=1}^{L} \sum_{x \in T_i} d(\boldsymbol{x}, \boldsymbol{v}_i) \,. \tag{35}$$

Centroid $\boldsymbol{v}_i$ lze určit dle vztahu

$$\boldsymbol{v}_i = \frac{1}{n_i} \sum_{\boldsymbol{x} \in T_i} \boldsymbol{x} \,, \tag{36}$$

kde n_i je počet vektorů ve shluku T_i.

[1]

5.1.1 MacQueenův algoritmus

MacQueenův algoritmus, někdy také uváděn jako *k-means*, je efektivním algoritmem pro tvorbu shluků. Zásadním problémem ve shlukové analýze je fakt, že existuje obrovské množství možností, jak rozdělit N vstupních vektorů do L neprázdných shluků. Například pro $L = 3$ a $N = 100$ existuje až 10^{47} možností. MacQueenův algoritmus je vhodným řešením z hlediska kvality rozložení shluků a také výpočetních nároků.

Nyní bude popsán základní algoritmus v několika krocích. Jsou dána následující označení:

$T_i(k)$ … množina vektorů i – tého shluku v k – tém kroku iterace,
$v_i(k)$ … centroid i – tého shluku v k – tém kroku iterace,
$J_i(k)$ … dílčí hodnota kritéria i – tého shluku v k – tém kroku iterace,
$n_i(k)$ … počet vektorů $\boldsymbol{x}$ ve shluku T_i v k – tém kroku iterace.

Postup algoritmu:

1. Je náhodně vybráno L počátečních centroidů $\boldsymbol{v}_1(1)$, $\boldsymbol{v}_2(1)$,…, $\boldsymbol{v}_L(1)$.
2. V k-té iteraci se rozdělí vektory trénovací množiny T do L shluků $T_1(k)$,…,$T_2(k)$ dle vztahu

$$\boldsymbol{x} \in T_j(k) \text{ jestliže } d(\boldsymbol{x}, \boldsymbol{v_j}(k)) < d(\boldsymbol{x}, \boldsymbol{v}_i(k)) \tag{37}$$

   pro všechny $i, j = 1,\ldots, L$, přičemž $i \neq j$. Tento vztah je postupně aplikován na všechny vektory v trénovací množině T.

3. Z výsledků bodu 2 je pro každý shluk vypočten nový centroid $\boldsymbol{v}_i(k+1)$ (kde $j = 1,\ldots, L$) tak, aby suma míry zkreslení všech vektorů množiny $T_j(k)$ vůči novému centroidu byla minimální. Centroid $\boldsymbol{v}_j(k+1)$ minimalizuje kritérium

$$J_j(k+1) = \sum_{\boldsymbol{x} \in T_j(k)} d\left(\boldsymbol{x}, \boldsymbol{v}_j(k+1)\right), \qquad j = 1, \ldots, L, \tag{38}$$

   a je vypočten dle vztahu

$$\boldsymbol{v}_j(k+1) = \frac{1}{n_j(k)} \sum_{\boldsymbol{x} \in T_j(k)} \boldsymbol{x}, \qquad j = 1, \ldots, L. \tag{39}$$

4. Pokud $\boldsymbol{v}_j(k+1) = \boldsymbol{v}_j(k)$ pro $j = 1,\ldots, L$ nebo je – li pokles celkového zkreslení

$$J(k) = \sum_{i=1}^{L} J_i(k), \tag{40}$$

   v k-té iteraci vzhledem k $J(k-1)$ pod předem definovaným prahem, je činnost algoritmu ukončena. Pokud těmto podmínkám není vyhověno, algoritmus pokračuje opět bodem 2 až do splnění některé z těchto dvou podmínek. Výsledkem na výstupu algoritmu je L centroidů $\boldsymbol{v}_i$ reprezentujících vektorový kvantizér. Vektory, které náleží jednotlivým shlukům T_j jsou společně s aplikovanou mírou zkreslení podkladem pro vymezení oblastí X_j, pro $j = 1,\ldots, L$. [1]

5.2 *k-NN* klasifikace

Klasifikátor *k-NN* (z angl. k nearest neighbours, zkr. *k-NN*) patří do skupiny učících se klasifikátorů a je možno jej použít online. Předpokladem pro použití tohoto klasifikátoru je apriorní existence trénovací množiny o *n* prvcích, získané většinou prostřednictvím shlukové analýzy (MacQueenův algoritmus). Základní princip klasifikace nově příchozího objektu, jímž může být např. vektor příznaků segmentů řeči, je založen na velikosti jeho vzdálenosti od *k* nejbližších objektů (vektorů), které náleží jednotlivým shlukům. Shluky zde představují určité konkrétní definované třídy, do kterých může být objekt přiřazen. Klasifikovaný objekt je přiřazen té třídě, která z celkových *k* naleznutých nejbližších objektů zaujímá jejich nejvyšší počet.
Nechť $\boldsymbol{y}$ je neznámý objekt (vektor) a $\boldsymbol{x}_i$ je objekt trénovací množiny, kde $i = 1,\ldots, n$. Princip funkce klasifikátoru je znázorněn následujícím pseudokódem (obrázek 14). [8]

```
BEGIN
    Zadej neznámý vektor y
    Zadej k, počet nejbližších sousedů,   1 ≤ k ≤ n
    Inicializuj i = 1

    DO UNTIL (nalezeno k-nejbližších sousedů)
        Vypočti vzdálenost od y do x_i
        IF (i ≤ k) THEN
            Zahrneme x_i do množiny k-nejbližších sousedů
        ELSE
            IF (x_i je blíže k y než jakýkoliv předchozí NN) THEN
                Vymaž nejvzdálenější z množiny k-NN
                Zahrň x_i do množiny k-NN
            END IF
        Zvětši i (i = i + 1)
    END DO UNTIL

    Urči majoritní třídu v množině k-NN (kde je nejvíce členů)
    Tam přiřaď y
END
```

Obrázek 14: Pseudokód *k-NN* klasifikátoru. [8]

5.3 Algoritmus *DTW*

Algoritmus *DTW* (z angl. dynamic time warping, zkr. DTW), neboli metoda dynamického borcení časové osy, je efektivním nástrojem pro klasifikaci izolovaně řečených slov a za použití určitých modifikací i řeči souvislé. Nutností je použití řečového detektoru. Jeho prostřednictvím je určen výběr v čase po sobě jdoucí sekvence segmentů řečového signálu. Z těchto jsou extrahovány vektory příznaků jak pro potřebu získání referenčních vzorových příznaků konkrétních slov (tzv. režim trénování), tak pro získání vektorů určených k porovnání s předem získanými referenčními příznaky každého slova (režim klasifikace). Zásadním problémem při rozpoznávání slov je fakt, že řečník nikdy neřekne totéž slovo v naprosto shodném časovém intervalu a doba trvání jednotlivých fonémů se při každé promluvě může významně lišit. Algoritmus *DTW* tento problém řeší pomocí nelineární časové normalizace. Je dán obraz testovaného slova

$$\boldsymbol{A} = \{\boldsymbol{a}(1), \boldsymbol{a}(2), \dots, \boldsymbol{a}(n), \dots, \boldsymbol{a}(I)\} \tag{41}$$

a obraz referenčního slova

$$\boldsymbol{B} = \{\boldsymbol{b}(1), \boldsymbol{b}(2), \dots, \boldsymbol{b}(m), \dots, \boldsymbol{b}(J)\}\ , \tag{42}$$

kde $\boldsymbol{a}(n)$ je n-tý vektor příznaků testovaného slova a $\boldsymbol{b}(m)$ je m-tý vektor příznaků slova referenčního. I a J je konečný počet vektorů příznaků. Algoritmus poté hledá v rovině (m,n) optimální cestu

$$m = \boldsymbol{\psi}(n)\ , \tag{43}$$

která minimalizuje funkci D celkové vzdálenosti mezi obrazy $\boldsymbol{A}$ a $\boldsymbol{B}$

$$D(\boldsymbol{A}, \boldsymbol{B}) = \sum_{n=1}^{I} \hat{d}[\boldsymbol{a}(n), \boldsymbol{b}(\boldsymbol{\psi}(n))]\ , \tag{44}$$

kde $\hat{d}[\boldsymbol{a}(n),\boldsymbol{b}(\boldsymbol{\psi}(n))]$ je lokální vzdálenost mezi n-tým vektorem příznaků testovaného slova a m-tým vektorem příznaků slova referenčního.

Stejně jako v případě vektorové kvantizace by měl být druh míry vzdálenosti volen v závislosti na typu používaných příznaků. Princip základního algoritmu *DTW* je dán následujícími kroky.

1. Je vytvořena matice $\boldsymbol{D}$ o I řádcích a J sloupcích, která je vyplněna vzájemnými lokálními vzdálenostmi $d(n,m) = d[\boldsymbol{a}(n),\boldsymbol{b}(m)]$.

2. Je vytvořena matice $\boldsymbol{G}$ částečných akumulovaných lokálních vzdáleností $g(n,m)$, o I+1 řádcích a J+1 sloupcích, kde nultý řádek a nultý sloupec jsou vyplněny nekonečnými hodnotami, s výjimkou pozice $g(0,0)$, která musí být inicializována nulovou hodnotou.

3. Je nutno vybrat typ metody omezení lokální cesty *DTW* . Existuje celkem sedm základních typů funkcí lokálních omezení, kde každý typ disponuje charakteristickými vlastnostmi strmosti a váhovou funkcí $\widehat{W}(k)$. Proměnná k je obecná časová proměnná a časové proměnné m a n jsou takovými funkcemi k, že platí

$$n = i(k)\,, \tag{45}$$
$$m = j(k)\,, \tag{46}$$

kde $k = 1, \ldots, K$, přičemž K je délka společné obecné časové osy pro porovnání obrazů $\boldsymbol{A}$ a $\boldsymbol{B}$.
V případě rozpoznávání izolovaných slov musí být přesně určeny počáteční a koncové hraniční body pro testovaný i referenční obraz dle podmínek

$$\begin{aligned} i(1) &= 1, \qquad j(1) = 1\,, \\ i(K) &= I, \qquad j(K) = J\,. \end{aligned} \tag{47}$$

Váhová funkce $\widehat{W}$ závisí na lokální cestě, existuje pět základních typů váhové funkce. Jako příklad je možno uvést jednu z těchto funkcí:
typ a) symetrická váhová funkce

$$\widehat{W}(k) = [i(k) - i(k-1)] + [j(k) - j(k-1)]\,, \tag{48}$$

přičemž $i(0) = j(0) = 0$.
Jednotlivé prvky matice $\boldsymbol{G}$ jsou rekurzivně vypočteny dle obecného vztahu

$$g[i(k), j(k)] = \min_{\{i(k), j(k)\}} \{g[i(k-1), j(k-1)] + d[i(k), j(k)]\widehat{W}(k)\}\,, \tag{49}$$

kde $k = 1, \ldots, K$, a $d[i(k), j(k)]$ je příslušná lokální vzdálenost z matice $\boldsymbol{D}$. Je li do vztahu (49) dosazena váhová funkce typu a), pak lze odvodit konkrétní vztah pro výpočet lokálního omezení typu I, který je dán vztahem

$$g(n,m) = min \begin{cases} g(n, m-1) + d(n,m) \\ g(n-1, m-1) + 2d(n,m)\,. \\ g(n-1, m) + d(n,m) \end{cases} \tag{50}$$

Konečnou normalizovanou vzdálenost mezi obrazy $\boldsymbol{A}$ a $\boldsymbol{B}$ lze vyčíslit dle vztahu

$$D(A,B) = \left[N(\widehat{W})\right]^{-1} g[i(K), j(K)] = \left[N(\widehat{W})\right]^{-1} g[I,J]\,, \tag{51}$$

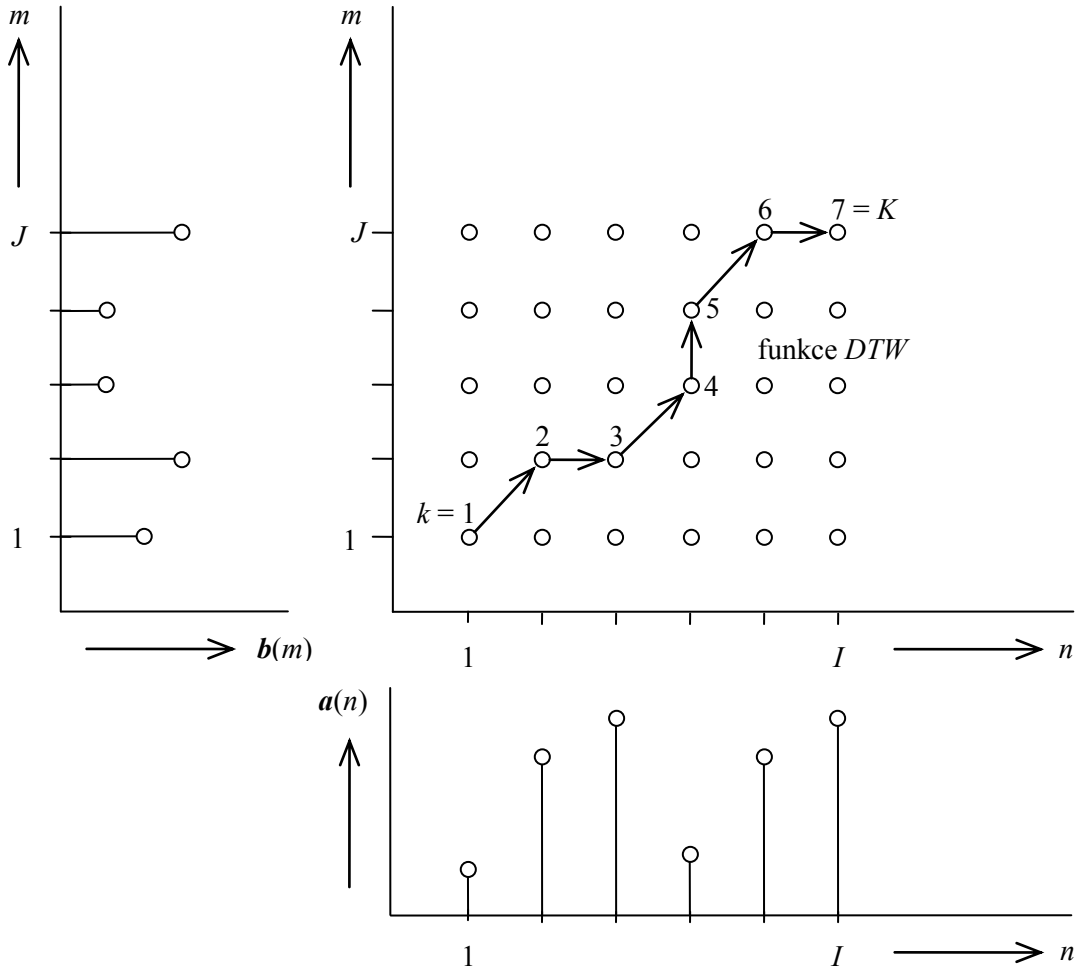
kde $N(\widehat{W})$ je normalizační faktor, který je odvozen z váhové funkce a je dán vztahem

$$N(\widehat{W}) = \sum_{k=1}^{K} \widehat{W}(k)\,. \tag{52}$$

V případě váhové funkce typu a) bude výsledný normalizační faktor určen vztahem

$$N(\widehat{W}) = I + J\,. \tag{53}$$

V případě nejvyšší shody obrazu $\boldsymbol{A}$ s obrazem $\boldsymbol{B}$ bude tedy celková vzdálenost mezi obrazy reprezentována hodnotou 0, v opačném případě pak hodnotou 1.
Na obrázku (obrázek 15) je znázorněn způsob nalezení optimální cesty funkcí *DTW*, kterou lze ovlivnit výběrem typu lokálního omezení.



Obrázek 15: Schematické znázornění porovnání testovaného a vzorového obrazu v rovině (*n*,*m*). [1]

Dodatečně je nutno poznamenat, že volba druhu míry vzdálenosti či zkreslení závisí na typu příznaků. Jsou li vektory příznaků tvořeny koeficienty *LPC*, je doporučeno použít tzv. nesymetrickou Itakurovu míru zkreslení dle vztahu

$$d(\boldsymbol{a}(n), \boldsymbol{b}(m)) = \log\left[\frac{\boldsymbol{b}^T(m)\boldsymbol{R}(n)\boldsymbol{b}(m)}{\boldsymbol{a}^T(n)\boldsymbol{R}(n)\boldsymbol{a}(n)}\right], \tag{54}$$

kde $\boldsymbol{R}(n)$ je autokorelační Töplitzova matice testovaného segmentu řádu $(Q + 1)\times(Q + 1)$, přičemž Q je řád lineárního prediktoru.

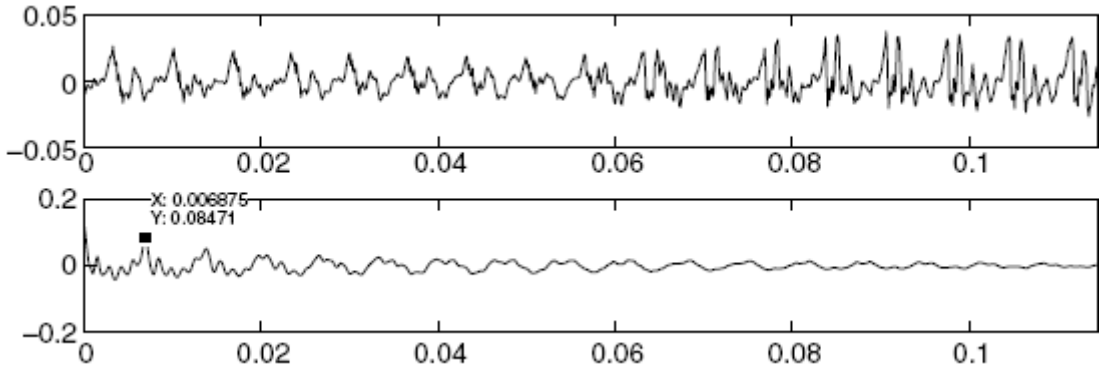
[1], [9]

6 Fonetická analýza řeči, dekódování řeči

6.1 Určení základního tónu řeči

6.1.1 Určení F_0 pomocí autokorelační funkce

Fonetická analýza se zabývá zejména určením základní frekvence, resp. hlasivkového tónu a formantových frekvencí. Jak již bylo řečeno, základní neboli fundamentální tón lidského hlasu lze získat výpočtem autokorelační funkce. Na obrázku (obrázek 16) je znázorněna autokorelační funkce $R(\tau)$ znělého segmentu řeči $s(t)$ odpovídající časové délce 12 ms. Informaci o hodnotě základní periody T_0 lze určit jako maximální postranní špičku funkce $R(\tau)$, neboli tzv. *LAG*, vzdálený od počátku souřadného systému. Protože autokorelační funkce je funkcí sudou, pro účely analýzy postačí její pravá polovina v kladné oblasti nezávislé proměnné.



Obrázek 16: Segment znělé promluvy (nahoře) a jeho autokorelační funkce (dole). [2]

Záměrně by hodnota fundamentální periody měla být hledána v intervalu, který odpovídá rozsahu lidského hlasu, tedy kmitočtům 50-400 Hz. Tímto se částečně zamezí možnosti chybného odečtu hodnot. Nevýhodou použití této funkce je riziko chybného odečtení hodnoty *LAG*u, způsobené vlivem formantové struktury, tedy přítomnost dalších vrcholů kolem prvního maxima. Tomuto lze předejít použitím modifikované autokorelační funkce (*MACF*), prostřednictvím filtrace řečového signálu pomocí *FIR* filtru typu dolní propust s dolní mezní frekvencí 900 Hz a následným zploštěním spektra.

[2], [1]

6.1.2 Určení F_0 pomocí kepstrální metody

Kepstrální metoda určení základního tónu vychází z diskrétní Fourierovy transformace a má za úkol od sebe oddělit vliv buzení (základního tónu hlasivek) a artikulačního ústrojí (formantových frekvencí). Protože prostá spektrální analýza k tomuto úkolu nestačí, je zapotřebí provést několik modifikací. Výpočet kepstra je dán následujícím vztahem

$$c(n) = IDFT\{log|DFT\{s(n)\}|^2\}, \tag{55}$$

kde $s(n)$ je diskrétní řečový signál vzniklý konvolucí dle vztahu (1). Platí-li, že

$$DFT\{s(n)\} = S(f) = G(f)H(f), \tag{56}$$

pak lze napsat, že

$$|DFT\{s(n)\}|^2 = |S(f)|^2 = |G(f)|^2|H(f)|^2. \tag{57}$$

Dalším krokem výpočtu je provedení inverzní diskrétní Fourierovy transformace, která patří mezi transformace lineární. Aby mohly být při zpětné Fourierově transformaci odděleny složky buzení a artikulačního systému, je zapotřebí provést logaritmizaci. Tím dojde k převedení operace součinu dvou členů na součet dvou logaritmů těchto členů, a inverzní Fourierova transformace se tak rozdělí na dvě samostatné části

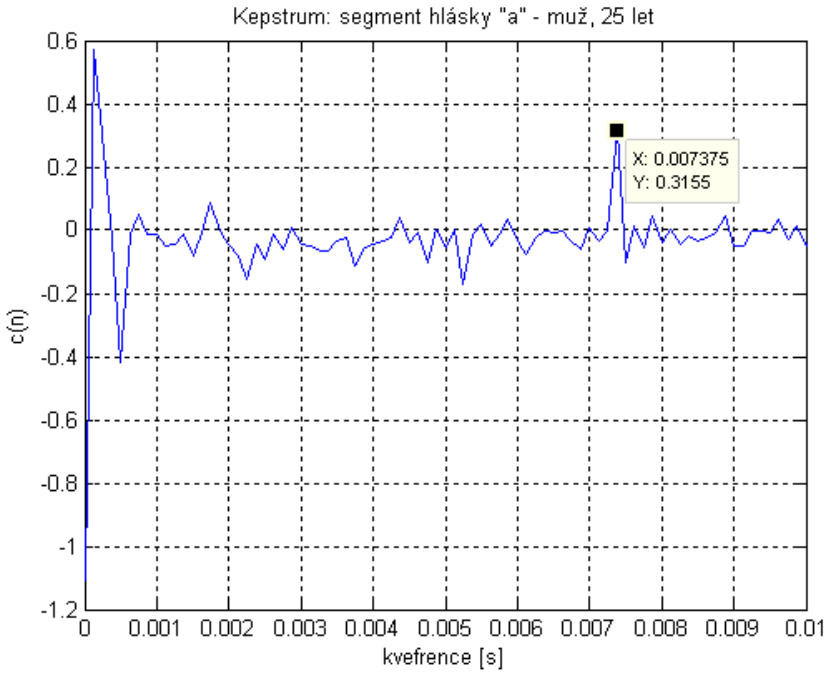
$$IDFT\{\log(|G(f)|^2|H(f)|^2)\} = IDFT\{\log|G(f)|^2\} + IDFT\{\log|H(f)|^2\}, \tag{58}$$

$$IDFT\{\log|G(f)|^2\} + IDFT\{\log|H(f)|^2\} = c_g(n) + c_h(n), \tag{59}$$

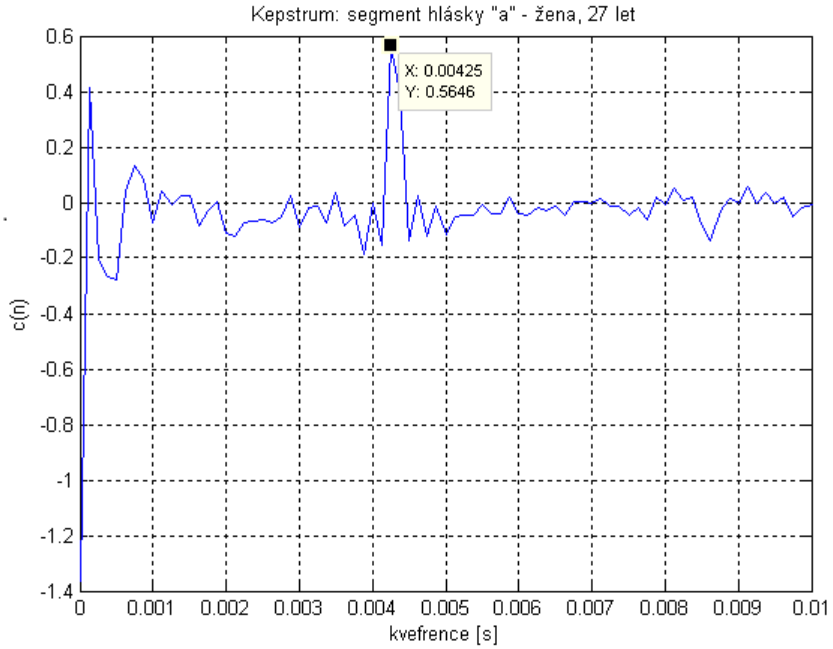
kde $c_g(n)$ jsou kepstrální koeficienty náležící buzení a $c_h(n)$ jsou kepstrální koeficienty náležící artikulačnímu traktu.

Takto transformovaný signál je převeden zpět do časové oblasti. Nejedná se ovšem o jeho zpětnou rekonstrukci, ale o promítnutí oddělených spektrálních vlastností hlasivek a artikulačního traktu do spektra period. Kepstrum lze tedy pokládat za spektrum spektra, o čemž svědčí i záměna písmen v jeho názvu. V tomto případě lze vodorovnou osu nazvat osou kvefrenční. Lze říci, že právě skloubení linearity Fourierovy transformace a nelinearity logaritmické funkce dělá z kepstra velmi důmyslný matematický nástroj. Na obrázcích níže jsou znázorněna výkonová kepstra 20 ms úseku hlásky „a“, znázorňující reálné výkonové kepstrální charakteristiky u muže ve věku 25 let (obrázek 17) a ženy, ve věku 27 let (obrázek 18). Jsou vyznačeny špičky, jenž odpovídají charakteru základní periody hlasivek. Po spočtení převrácené hodnoty kvefrence (periody) daných špiček vychází u muže základní tón hlasivek roven přibližně 135 Hz, u ženy pak 235 Hz. Všechny ostatní kvefrence, směřující od základní kvefrence napravo, směrem k počátku souřadného systému náleží systému artikulačnímu a lze je po úpravách dále použít např. pro určování charakteristik formantů.

[4],[3],[2]



Obrázek 17: Reálné výkonové kepstrum - muž.



Obrázek 18: Reálné výkonové kepstrum - žena.

6.2 Akusticko-fonetické dekódování řeči

Úkolem metod akusticko-fonetického dekódování je klasifikovat detekované segmenty řeči a v konečné fázi jim přiřadit značku znělosti, konkrétní samohlásky, okluzívy, apod. Tyto metody lze z hlediska přístupů rozdělit na tři základní, a to na:

- přístupy založené na heuristických pravidlech
- přístupy založené na vzdálenosti
- přístupy pravděpodobnostní (*HMM*)

Snahou heuristického přístupu je aplikovat znalosti expertů (fonetiků) do tzv. expertních systémů, čili systémů schopných rozhodování a klasifikace na základě vžitých zkušeností daných jedinců. Tyto znalosti bývají vyjádřeny formou tzv. produkčních pravidel ve tvaru:

JESTLIŽE podmínka **PAK** závěr.

Vstupními parametry těchto pravidel mohou být formantové frekvence, frekvence základního tónu řeči, apod. Velkou výhodou tohoto přístupu je možnost vyvíjet systém na základě subjektivních pozorování a intuici konstruktéra.

Přístupy založené na vzdálenosti fungují na základě porovnávání segmentů řeči se souborem referenčních segmentů, u kterých je známo přesné přiřazení. Předpokladem pro tuto metodu je nalezení změn znělosti, intenzity či ostré spektrální změny, způsobené různými konfiguracemi artikulačního ústrojí. Jsou-li si dva segmenty svými parametry dosti podobné, lze s vysokou pravděpodobností říci, že se jedná o segmenty stejného druhu. [1]

6.2.1 Fonetická transkripce

Fonetická transkripce přiřazuje zvukům mluvené řeči přesnou a nedvojznačnou textovou formu. Nejobecnější mezinárodní fonetická abeceda *IPA* (z angl. International Phonetic Alphabet) (obrázek 19) umožňuje přepsat promluvu v libovolném jazyce prostřednictvím systému symbolů, náležících např. jednotlivým fonémům nebo alofónům (tj. fonémům, u kterých vlivem některé okolní hlásky dojde ke změně jejich artikulace). Výhodou přepisu zvukové podoby českého jazyka je možnost použití větší části znaků obyčejné české abecedy, s případným doplněním některých pomocných symbolů. V počítačovém zpracování řeči se využívá fonetické abecedy *SAMPA* (z angl. Speech Assesment Methods Alphabet) (obrázek 20), která je postavena na konverzi abecedy *IPA* do sedmibitové *ASCII* sady. Je třeba dodat, že hlavním polem využití fonetické transkripce je v současnosti rozpoznání řeči pomocí *HMM*, kde jsou slova modelována řetězci modelů fonémů. Dále pak může být využita při hlasové syntéze z psaného textu. [1], [10]

THE INTERNATIONAL PHONETIC ALPHABET (revised to 1993)

CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap				ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

Obrázek 19: Ukázka části abecedy IPA pro pulmonální souhlásky. [12]

Long vowels		Short vowels		Consonants	
SAMPA	IPA	SAMPA	IPA	SAMPA	IPA
a:	aː	a	a	b	b
e:	eː	E	ɛ	d	d
i:	iː	I	ɪ	g	g
o:	oː	O	ɔ	l	l
u:	uː	U	ʊ	x	x
E:	ɛː	Y	ʏ	r	ɾ, ʁ
2:	øː	9	œ	m	m
y:	yː	@	ə	n	n
		6	ɐ	s	s

Obrázek 20: Ukázka konverze symbolů IPA do SAMPA pro některé souhlásky. [11]

7 Praktická realizace systému pro detekci řeči

Vlastní návrh offline systému pro rozpoznání řeči sestává z intenzitního detektoru, vyhodnocovače základní frekvence hlasu F_0 a znělosti, vyhodnocovače osoby mluvčího (žena, muž, dítě), rozpoznávače izolovaných slov a rozpoznávače samohlásek *a, e, i, o, u*.

Uživatelské prostředí programu je uděláno jednoduše, interaktivní formou, pomocí grafických prvků a vyplnitelných textových polí, prostřednictvím výpočetního programu MATLAB (verze R2010a). Program je fyzicky obsažen v soborech *GUI_ASR.m* a *GUI_ASR.fig*.

7.1 Popis obsluhy uživatelského prostředí

Uživatelské prostředí systému (obrázek 21) umožňuje načtení jednokanálového zvukového záznamu ve formátu **.wav* (tl. *Načíst*, vpravo nahoře), popřípadě pořízení vlastní nahrávky (tl. *Nahrát*, vpravo nahoře, pod tl. *Načíst*) téhož formátu, kde délka nahrávky je volitelná dle libosti uživatele. Vzorkovací frekvence nahrávky musí být 16 kHz, v případě pořízení nahrávky v uživatelském prostředí je vzorkovací kmitočet již přednastaven. Po pořízení nahrávky je možno ji uložit na pevný disk (tl. *Uložit wav*, napravo od tl. *Nahrát*) nebo použít pro okamžitou analýzu.

Obecné údaje o délce nahrávky a vzorkovacím kmitočtu jsou po načtení řečového signálu zobrazeny v ovládacím panelu zvaném *Parametry nahrávky*, nacházejícím se vpravo nahoře, bezprostředně pod tlačítky *Nahrát* a *Uložit wav*, přičemž nahrávku lze také přehrát pomocí tlačítka *PŘEHRÁT*. Pod panelem *Zvukový záznam* je umístěn panel *Nastavení detektoru*, který slouží k nastavení vstupních parametrů detektoru řeči, a zahájení detekce tlačítkem *DETEKCE*.

Průběh amplitudy řečového signálu v čase je v grafu po načtení a detekci znázorněn modře, výsledky detekce červeně, znázornění znělosti zeleně (neznělá část slova je reprezentována hodnotou 0, znělá pak hodnotou 1) a průběh F_0 černě (viz příloha I). Průběhy lze mezi sebou kombinovat či přepínat (panel *Zobrazit* vlevo dole) v případě potřeby lze signál zvětšovat, zmenšovat či posunovat. K tomuto účelu slouží panel nástrojů, umístěný v levém horním rohu.

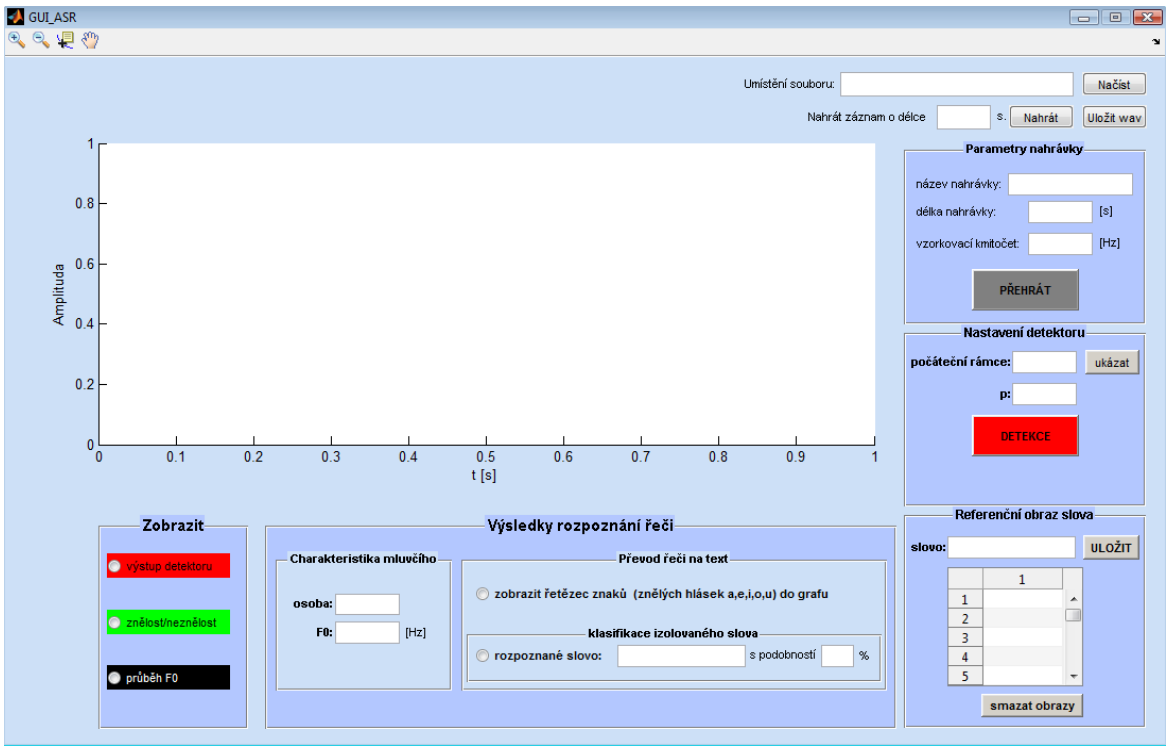
Po provedení detekce jsou v panelu *Výsledky rozpoznání řeči* (vpravo od panelu *Zobrazit*) automaticky vyplněny informace o druhu osoby a příslušné hodnotě F_0 (panel *Charakteristika mluvčího*).

Taktéž lze přímo do průběhu řečového signálu zobrazit textový výpis samohlásek *a*, *e*, *i*, *o*, *u*, jsou li ve slově obsaženy, pomocí tlačítka *zobrazit řetězec znaků* v panelu *Výsledky rozpoznání řeči* a subpanelu *Převod řeči na text*.

V případě rozpoznávání izolovaných slov je třeba nejprve nahrát pro každé slovo vzorovou promluvu, provést její detekci a do textového pole v panelu *Referenční obraz slova* (pod panelem *Nastavení detektoru*) napsat název daného slova, které se uloží do tabulky slov. V případě nekvalitní detekce lze promluvu či nastavení detekce opakovat až do dosažení optimálního výsledku. Poté je třeba název daného slova a tím i sekvenci *LPC* příznaků uložit prostřednictvím tlačítka *ULOŽIT,* které se nachází taktéž v panelu *Referenční obraz slova*. Poté může být provedeno rozpoznání testované promluvy. Nejprve je třeba testovací promluvu v nahrávce detekovat a poté, v panelu *Výsledky rozpoznání řeči* a subpanelu *klasifikace izolovaného slova*

zobrazit rozpoznané slovo prostřednictvím tlačítka *rozpoznané slovo*. Výstupem je výpis rozpoznaného slova do textového pole, nacházejícím se v témže subpanelu a procentuální shoda s nejvíce podobnou referencí. V opačném případě, kdy slovo rozpoznáno není, se do téhož textového pole vypíše text *nerozpoznáno*. Referenční obrazy slov lze v případě potřeby smazat tlačítkem *smazat obrazy*.

V případě neprovedení jakéhokoliv předcházejícího potřebného kroku k provedení detekce či rozpoznání dojde k upozornění formou okna výstrahy.



Obrázek 21: Uživatelské prostředí systému pro detekci řeči.

7.2 Popis implementovaných funkcí pro zpracování řeč. signálu

Nyní budou detailněji vysvětleny funkce a proměnné, které se přímo podílí na zpracování řečového signálu. Méně rozsáhlé funkce a proměnné mající spíše doplňující význam pro chod programu (nahrávání signálu, zobrazení grafů, výstražná sdělení apod.) budou popsány formou části kódu, resp. budou pouze slovně zmíněny. Zásadní algoritmy budou znázorněny pomocí vývojových diagramů.

7.2.1 Pořízení, předzpracování signálu a úprava dat

Jak bylo zmíněno v popisu obsluhy, nahrávku lze získat dvěma způsoby, a sice načtením či nahráním. Tlačítko *Načíst* volá funkci *nacti_signal*, která spouští okno pro načtení nahrávky. Vstupními proměnnými funkce jsou *detekovano*, *nacti* a *nahraj*, které jsou při načtení programu

inicializovány výchozí hodnotou 0, přičemž každým spuštěním funkce *nacti_signal* jsou proměnné *detekovano* a *nahraj* uvedeny do hodnoty 0 a proměnná *nacti* do hodnoty 1. Tyto proměnné slouží jako pomocné a jsou zapojeny do podmínek pro spouštění dalších funkcí, kde je vyžadována posloupnost úkonů nebo které potřebují mít pro svou činnost načteny aktuální data (např. detektor řeči). Výstupními proměnnými jsou pak aktualizované *detekovano*, *nacti*, *nahraj* a dodatečně *adresa* a *nazevwav*, reprezentující cestu ke zvukovému soboru a jeho název, sloužící k načtení dat z nahrávky. Jelikož jsou tyto proměnné vedeny jako globální, není třeba v kódu Matlabu formálně definovat vstupní a výstupní argumenty funkce.

Implementace kódu funkce *nacti_signal*

```
function nacti_signal(hObject,eventdata,handles)
global  detekovano adresa nazevwav nacti nahraj;
detekovano=0;
nacti=1;
nahraj=0;
[nazev_wav cesta_wav] = uigetfile({'*.wav'}, 'Vyberte soubor');
nazevwav=strcat(nazev_wav);%nazev souboru jako text
adresa = strcat (cesta_wav, nazev_wav);%kompletni adresa k souboru
% ----Zde je volána funkce otevri_wav
otevri_wav(handles);
```

U funkce *nahraj_signal* spouštěné tlačítkem *Nahrát* jsou vstupní proměnné tytéž, doplněny o lokální proměnnou *doba_nahr*, reprezentující uživatelem zadanou délku trvání nahrávky do textového pole. Voláním této funkce je proměnné *nahraj* udělena hodnota 1, proměnným *detekovano* a *nacti* pak hodnota 0. Na základě hodnoty proměnné *doba_nahr* a pevném vzorkovacím kmitočtu 16 kHz je pomocí funkce *wavrecord* pořízena nahrávka ve formátu *wav*. Výstupem funkce *nahraj_signal* je sloupcový vektor nahrávky *DATA* a aktualizované proměnné *detekovano*, *nacti*, *nahraj*, *adresa* a *nazevwav*. Protože nová nahrávka není uložena trvale na pevném disku ale pouze v paměti Matlabu, je proměnná typu *char adresa* počátečně nastavena na *Neuloženo* a proměnná *nazevwav* je prázdným polem znaků.

Implementace kódu funkce *nahraj_signal*

```
function nahraj_signal(hObject,eventdata,handles)
global detekovano DATA nahraj nacti adresa nazevwav;
adresa='Neuloženo';
nazevwav=[];
detekovano=0;
doba_nahr=str2double(get(handles.trvani_nahraj,'String'));
if isnan(doba_nahr) %osetreni vstupu
   warndlg('Zadejte dobu nahrávání v sekundách.','Zadejte délku záznamu!')
elseif doba_nahr<=0
   warndlg('Délka záznamu musí být větší než 0 sekund','Zadejte délku záznamu!')
else
nahraj=1;
nacti=0;
DATA=wavrecord(16000*doba_nahr,16000);%nahraj nahravku s Fvz 16000 kHz
```

```
otevri_wav(handles);
end
```

Každým zmáčknutím tlačítek *Načíst* a *Nahrát* je současně s voláním funkcí *nacti_signal* a *nahraj_signal* také inkrementována globální proměnná *obraz_poradi*, reprezentující pořadí referenčního slova ve slovníku (tabulce) pro referenční obrazy sekvence *LPC* příznaků. Byla-li nahrávka řeči pořízena vícekrát bez uložení příznaků (např. vlivem nekvalitního nastavení detektoru či omylu při výběru uložené nahrávky), je dodatečně inkrementována globální proměnná *posun_tabulka*, která zajišťuje postupné plnění tabulky bez mezer v řádcích.

Implementace kódu pro ošetření plnění slovníku

```
obraz_poradi=obraz_poradi+1;

if obraz_poradi>1
  if isempty(slovo_tabulka{obraz_poradi-posun_tabulka-1,1})
    posun_tabulka=(posun_tabulka+1);
  else
    posun_tabulka=posun_tabulka;
  end
elseif obraz_poradi==1
   posun_tabulka=posun_tabulka;
end
```

Případné uložení pořízené nahrávky na pevný disk je umožněno pomocí funkce *uloz_nahravku_wav* (spouštěna tlačítkem *Uložit wav*), která uloží původní neupravený signál reprezentovaný vektorem *DATA*. V případě prázdné proměnné *DATA* je na tuto skutečnost obsluhující osoba upozorněna dialogovým oknem. Je li nahrávka k dispozici, je uložena pomocí funkce *wavwrite* na příslušné místo (proměnná *cesta_wav*) a pod vybraným názvem (proměnná *nazev_wav*).

Implementace kódu funkce *uloz_nahravku_wav*

```
function uloz_nahravku_wav(hObject,eventdata,handles)
global DATA;

if isempty(DATA)
   warndlg('Nahrávka není k dispozici. Načtěte nebo nahrajte zvukový záznam.','Zvuková data
nejsou k dispozici!')
else
[nazev_wav, cesta_wav]=uiputfile('*.wav', 'Uložení souboru wav');
wavwrite(DATA, 16000,[cesta_wav nazev_wav])
adresa = strcat (cesta_wav, nazev_wav);
set(handles.umisteni_nacist,'String',adresa);
set(handles.nahravka_nazev,'String',nazev_wav);
end
```

Při pořízení nahrávky skrze funkci *nahraj_signal* je datový vektor *DATA* k dispozici ihned po ukončení nahrávání a předzpracování (normalizace atd.), zatímco v rámci načtení nahrávky skrze funkci *nacti_signal* jsou nejprve k dispozici pouze údaje o lokaci a názvu nahrávky.

V případě, že by obsluhující osoba stornovala načtení nahrávky, musí být další zpracování signálu ukončeno. To je ošetřeno ve funkci *otevri_wav* (volána v posledním řádku funkcí *nahraj_signal* a *nacti_signal*), která zajišťuje prvotní předzpracování řečového signálu, a sice ustředění, preemfázi a normalizaci amplitudy. Vstupními proměnnými jsou *DATA*, *nacti*, *nahraj*, *adresa* a *nazevwav*. V případě načtení dat z pevného disku se proměnná *DATA* vyplní hodnotami teprve na základě existence získané adresy lokace a názvu nahrávky z proměnných *adresa* a *nazev_wav*. Je li k dispozici sloupcový vektor *DATA*, je nejprve proveden výpočet střední hodnoty signálu, která je od něj posléze odečtena. Takto upravený signál je uložen do globální proměnné *U_DATA*. V dalším kroku je provedena preemfáze, čili odečtení *k* násobku (hodnota *k* je rovna 0.95) minulé hodnoty amplitudy od hodnoty aktuální a proměnná *U_DATA* je těmito hodnotami aktualizována. V posledním kroku je řečový signál amplitudově normalizován. Důležitými výstupními proměnnými jsou poté *U_DATA*, konstanta *pocet_vz_1segmentu* reprezentující číslo udávající počet vzorků jednoho segmentu řeči, konstanta *cas_delka_segmentu*, udávající časovou délku segmentu (32 ms) a *nvz*, tedy celkový počet vorků v nahrávce. Po načtení a zpracování je signál *U_DATA* automaticky vykreslen do grafu.

Implementace části kódu funkce *otevri_wav*

```
uDATA=(1/nvz)*sum(DATA(1:nvz,1)); %velikost str. hodnoty suroveho signalu
U_DATA = DATA - uDATA; %zprumerovany signal
% Preemfaze - zvyrazneni vyssich frekvenci v signalu
k=0.95;
  U_DATA(1,1)=U_DATA(1,1);
  U_DATA(2:nvz,1)=U_DATA(2:nvz,1)-k*U_DATA((2:nvz)-1,1);
%-------------------------------------------------------------%
U_DATA = U_DATA/max(abs(U_DATA));% normalizovany datovy vektor (max. amplituda vzdy
1)
cas_delka_segmentu =32*10^-3; % pozadujeme segment o delce 32 ms
pocet_vz_1segmentu = floor(Fvz*cas_delka_segmentu); % dany pocet vzorku 1 segmentu
```

Načtený signál *U_DATA* lze přehrát tlačítkem *PŘEHRÁT*, volajícím funkci *prehraj*, pomocí příkazu *wavplay*. Globální proměnná *Fvz* reprezentuje hodnotu vzorkovacího kmitočtu. Není li datový vektor *U_DATA* naplněn hodnotami, dojde ke spuštění dialogového okna s výstrahou.

Implementace části kódu funkce *prehraj*

```
if isempty(U_DATA)==0
   wavplay(U_DATA,Fvz);
elseif isempty(U_DATA)==1
   warndlg('Nejprve načtěte/nahrajte zvukový záznam.','Nebyla pořízen zvukový záznam!');
end
```

Před provedením detekce je třeba určit počáteční počet segmentů, ze kterých je počítána průměrná hodnota intenzity pro určení inicializačního prahu detekce. Tyto segmenty jsou získávány z pauzy před počátkem promluvy. Aby mohl být znázorněn časový úsek, v němž jsou počáteční segmenty obsaženy, je tlačítkem *ukázat*, jenž je umístěno v panelu *Nastavení detektoru*, volána funkce *poc_ram_ukaz*, která na základě globálních proměnných *Fvz* (vzorkovací frekvence),

pocet_vz_1segmentu a hodnoty počtu segmentů vložených uživatelem do textového pole *poc_ramce* vypočte horní hranici časového úseku, která je poté graficky znázorněna žlutou svislou čarou.

Implementace části kódu funkce *poc_ram_ukaz*

```
ukazat_pocatecni_ramce=str2double(get(handles.poc_ramce,'String'));
if isnan(ukazat_pocatecni_ramce)
warndlg('Vložte celé nezáporné číslo větší než 0','Počet počátečních rámců není číslo!');
elseif ukazat_pocatecni_ramce <=0
warndlg('Vložte celé nezáporné číslo větší než 0','Počet počátečních rámců byl zvolen chybně!');
elseif isempty(U_DATA)
warndlg('Nejprve načtěte/nahrajte zvukový záznam!','Zvukový záznam není k dispozici');
else
plot(handles.graf,(ukazat_pocatecni_ramce*((pocet_vz_1segmentu/2)/Fvz)),-1:0.01:1,'y');
end
```

Je li provedena detekce řečového signálu (globální proměnná *detekovano* nabude hodnoty 1), je možno uložit získanou matici *LPC* příznaků *a_t* do buňky pole globální proměnné *a_tR* (pole referenčních příznaků), na pozici určenou aktuálními hodnotami vstupních globálních proměnných *obraz_poradi* a *posun_tabulka*. Před provedením uložení je nutno do textového pole *zadat_slovo_DTW* napsat název promluvy, aby mohl algoritmus klasifikace izolovaných slov dodatečně přiřadit ke každé buňce *LPC* příznaků odpovídající název rozpoznávaného slova. Funkce *obraz_ulozit* je volána tlačítkem *ULOŽIT* v panelu *Referenční obraz slova*, proti uložení prázdného řetězce znaků názvu slova je ošetřena.

Implementace části kódu funkce *obraz_ulozit*

```
if detekovano==1
slovo_tabulka{obraz_poradi-posun_tabulka,1}=get(handles.zadat_slovo_DTW,'String');
   if isempty(slovo_tabulka{obraz_poradi-posun_tabulka,1})== 0
   set(handles.slovnik_tabulka,'data',slovo_tabulka,'ColumnName',1);
   a_tR{1,obraz_poradi-posun_tabulka}=a_t;
   elseif isempty(slovo_tabulka{obraz_poradi-posun_tabulka,1})== 1
   warndlg('Napište název slova.',...
   'Nebyl vložen název slova!');
   end
elseif detekovano == 0
warndlg('Nejprve nahrajte řečový signál a poté proveďte detekci promluvy.', 'Proveďte potřebné
kroky!');
end
```

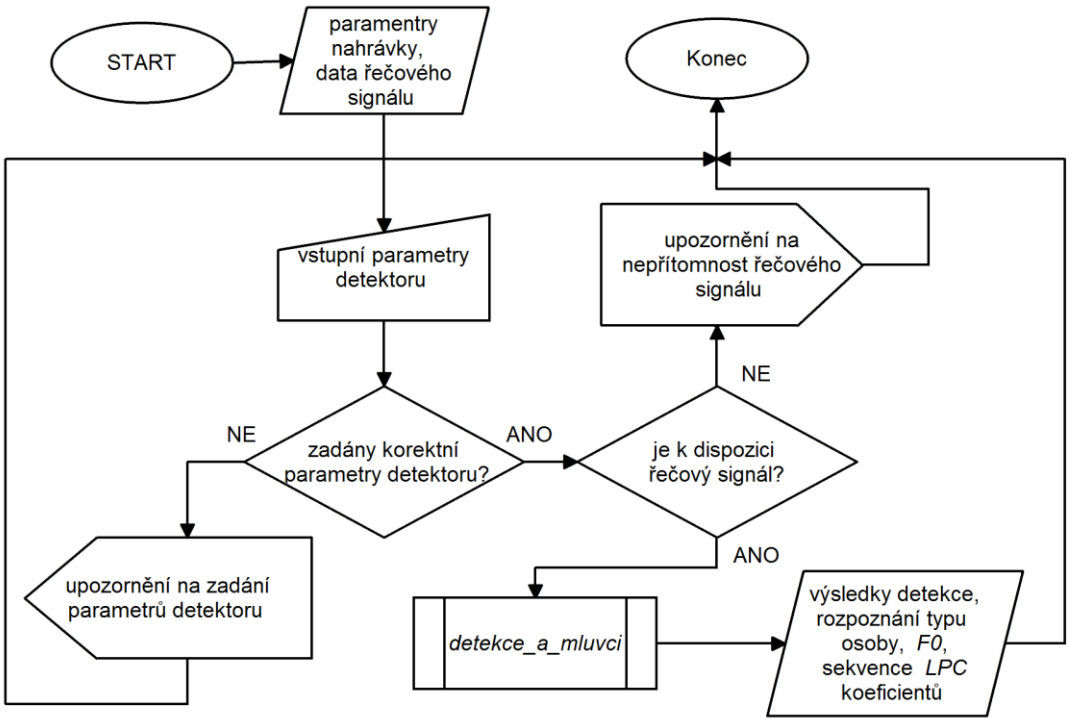
Získané referenční obrazy lze smazat prostřednictvím tlačítka *smazat obrazy*, které volá funkci *smazat_obrazy_LPC*. Proměnná *obraz_poradi* je nastavena na hodnotu 1 a *posun_tabulka* na výchozí hodnotu 0. Proměnná *a_tR* obsahující sekvence referenčních *LPC* příznaků je vymazána, pole *slovo_tabulka* je vyplněno prázdnými znaky, jimiž je tabulka slov v panelu *Referenční obraz slova* aktualizována.

Implementace části kódu funkce *smazat_obrazy_LPC*

```
obraz_poradi=1;
posun_tabulka=0;
clear a_tR;
for i=1:20
slovo_tabulka{i,1}='';
end
set(handles.slovnik_tabulka,'data',slovo_tabulka,'ColumnName',1);
```

7.2.2 Hlavní zpracování signálu

Mezi hlavní zpracování signálu je zahrnuta segmentace a detekce řeči, klasifikace znělosti segmentů spolu s výpočtem F_0 a určením typu mluvčího (muž, žena, dítě), dále je proveden výpočet koeficientů *LPC*. Výpočty jsou provedeny prostřednictvím spuštění funkce *detekce_a_mluvci* (tlačítko *DETEKCE*), jejíž jazykově popsaný princip spuštění včetně znázornění ošetření vstupních proměnných je na obrázku níže (obrázek 22).



Obrázek 22: Principielní schéma ošetření funkce *detekce_a_mluvci*.

Nyní bude detailněji popsána funkce *detekce_a_mluvci*. Důležité vstupní a výstupní proměnné jsou vypsány v tabulce (tabulka 4).

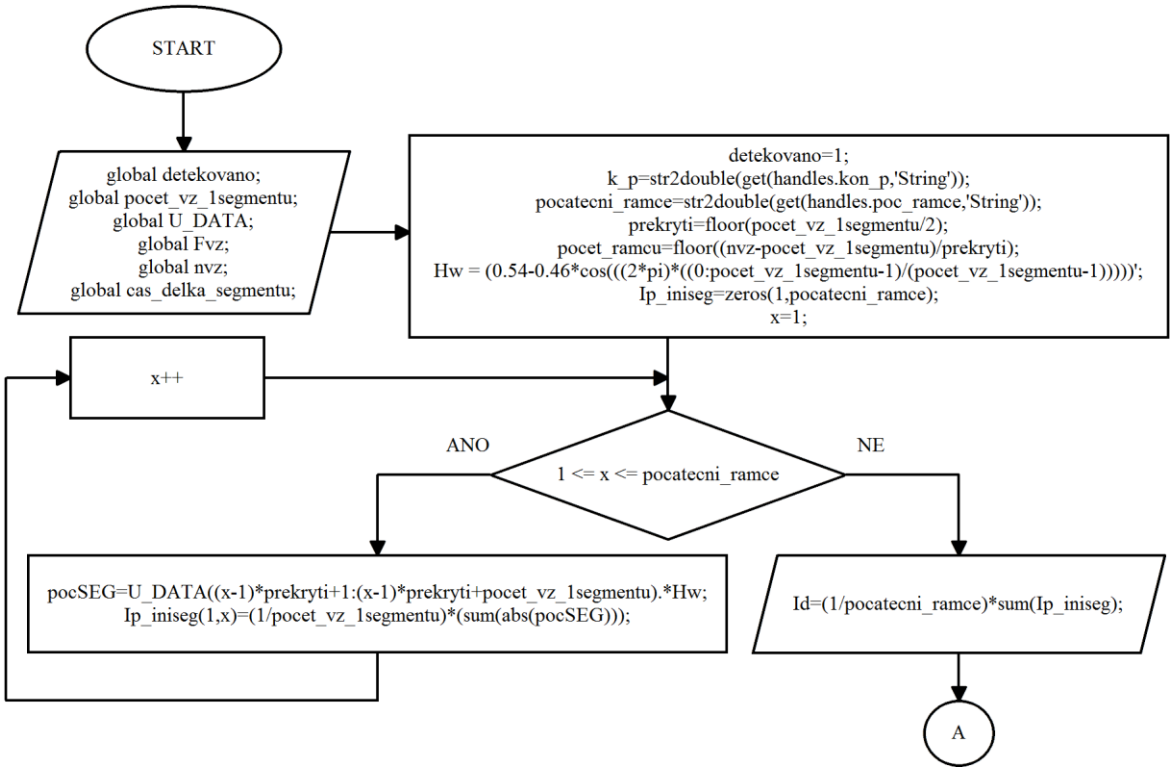
Tabulka 4: Vstupní a výstupní proměnné funkce *detekce_a_mluvci*.

Vstupní proměnná	Výstupní proměnná
global *detekovano*	global *detekovano*
global *U_DATA*	global *pocatecni_ramce*
global *Fvz*	global *prekryti*
global *nvz*	global *pocet_ramcu*
global *pocet_vz_1segmentu*	global *SEGMENTY*
global *cas_delka_segmentu*	global *DET_SEG*
	global *PAUZA*
	global *ZC*
	global *ZNELE_SEGMENTY*
	global *a_t*
	global *Htrakt*
	global *LT2*
	global *delka_filtru*

Ke spuštění funkce *detekce_a_mluvci* je potřeba, aby proměnná *U_DATA* byla naplněna hodnotami amplitudy řečového signálu, taktéž je nutno mít informaci o vzorkovacím kmitočtu (proměnná *Fvz*), počtu vzorků jednoho segmentu (proměnná *pocet_vz_1segmentu*) a z těchto dvou údajů plynoucí výsledek časové délky jednoho segmentu (proměnná *cas_delka_segmentu*). Dalším nutným údajem jsou inicializační parametry intenzitního detektoru, a sice počet inicializačních segmentů (proměnná *pocatecni_ramce*) a konstanta *p* detektoru (proměnná *k_p*). Proměnná *pocatecni_ramce* a proměnná *k_p* jsou definovány uvnitř funkce a načítány z vyplnitelných textových polí *poc_ramce* a *kon_p* z panelu *Nastavení detektoru*.

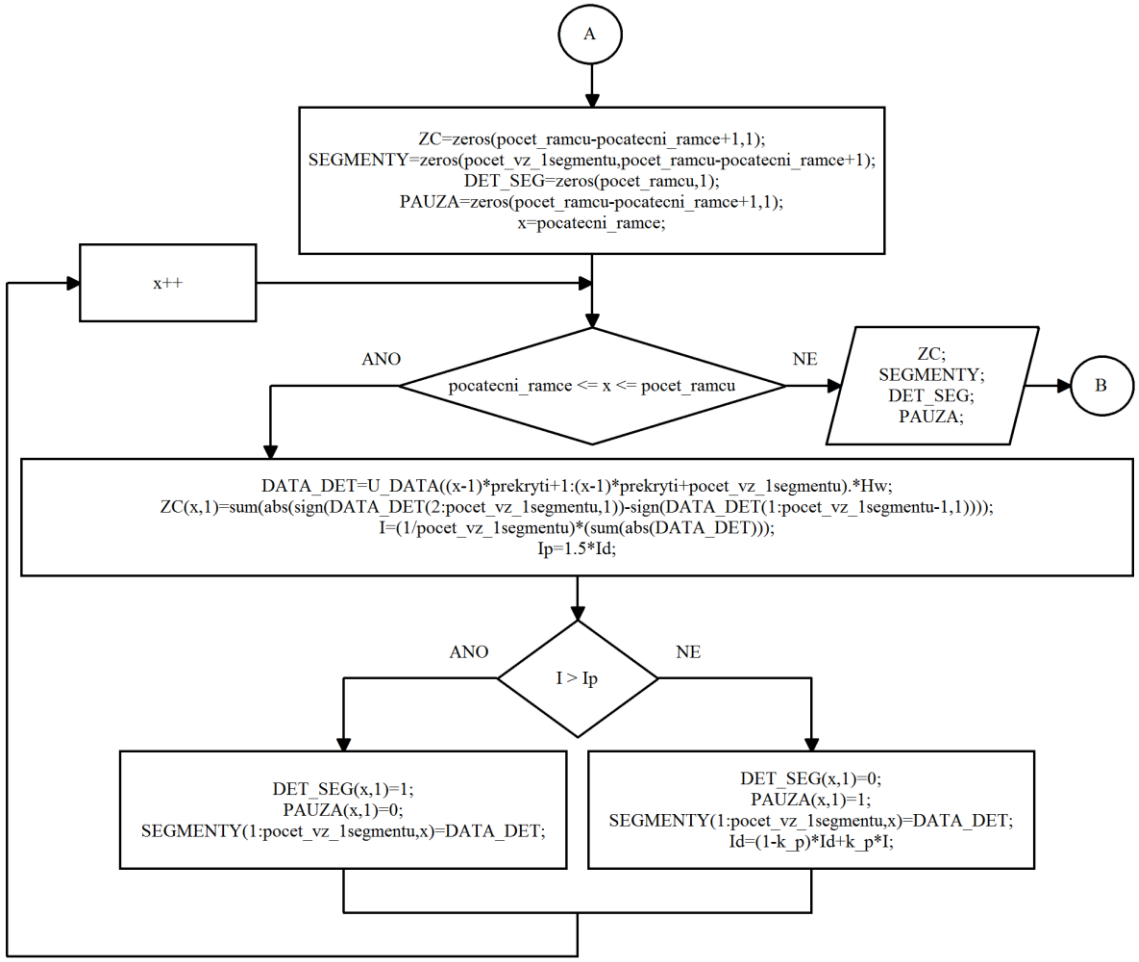
Jsou li načtena vstupní data, dojde nejprve k definování počtu vzorků určujících poloviční překrytí segmentů (proměnná *prekryti*) a k výpočtu celkového počtu překrývajících se segmentů (proměnná *pocet_ramcu*). Dále je definováno Hammingovo okno o délce počtu vzorků jednoho segmentu a jsou předdefinovány matice a vektory proměnných určené k ukládání výsledků detekce (tyto budou dále zmíněny).

Inicializace detektoru (obrázek 23) je zahájena cyklickým výpočtem normalizované intenzity signálu od prvního segmentu do celkového počtu inicializačních segmentů určeného proměnnou *pocatecni_ramce* Jednotlivé vektory počátečních segmentů (proměnná *pocSEG*) jsou extrahovány z vektoru *U_DATA* a váženy Hammingovým oknem *Hw*. Aktuální hodnoty intenzity těchto segmentů jsou cyklicky ukládány do řádkového vektoru *Ip_iniseg*. Po ukončení cyklu je vypočtena průměrná hladina intenzity přes všechny inicializační segmenty a uložena do lokální proměnné *Id*, která je podkladem pro určení detekčního prahu *Ip*.



Obrázek 23: Vývojový diagram inicializace intenzitního detektoru.

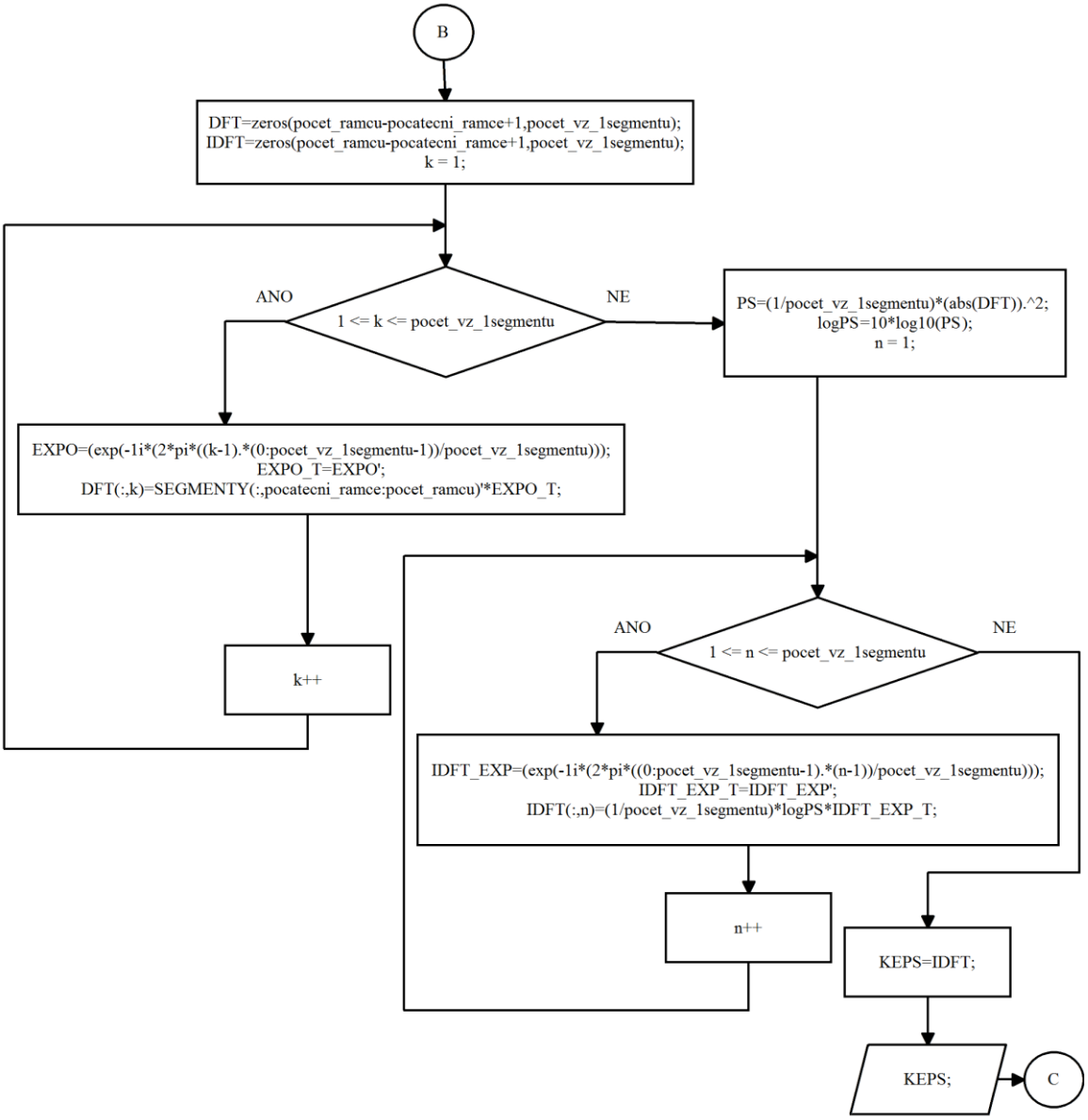
Po provedení inicializace může být provedena detekce přítomnosti řeči v signálu, na základě porovnání intenzity testovaných segmentů s prahovou hodnotou intenzity *Ip*, určenou dle vztahu (28). Obdobně jako v případě inicializace jsou cyklicky extrahovány segmenty z vektoru *U_DATA*, tentokráte do proměnné *DATA_DET*. Cyklus je inicializován pro segmenty náležící intervalu od konce inicializace (proměnná *pocatecni_ramce*) do celkového počtu segmentů (*pocet_ramcu*). Dále jsou pro každý testovaný segment vypočteny hodnoty počtu průchodu signálu nulou *ZC* a intenzita *I*. Aktualizace prahové hodnoty *Ip* je provedena v případě, že intenzita aktuálního segmentu je menší, než li prahová. V tomto případě je segment řečově neaktivní a do proměnné *DET_SEG*, která slouží jako registr detekovaných segmentů, je na aktuální pozici zapsána hodnota 0 a do proměnné *PAUZA* (registr časové pauzy v promluvě) je zapsána hodnota 1. V případě detekce řečové aktivity je do registru *DET_SEG* zapsána hodnota 1 a do registru PAUZA hodnota 0. Jednotlivé segmenty se ukládají coby sloupcové vektory do matice *SEGMENTY*, určené k dalšímu zpracování (obrázek 24).



Obrázek 24: Vývojový diagram průběhu detekce intenzitním detektorem.

Po naplnění matice *SEGMENTY* a ukončení činnosti detekce je provedena diskrétní Fourierova transformace jednotlivých segmentů za účelem získání krátkodobého odhadu jejich amplitudového spektra (proměnná *DFT*). Hodnoty výkonového spektra jednotlivých segmentů (proměnná *PS*) jsou poté logaritmovány a je provedena jejich inverzní Fourierova transformace. Výsledkem je matice vektorů kepstrálních koeficientů jednotlivých segmentů (proměnná *KEPS*).

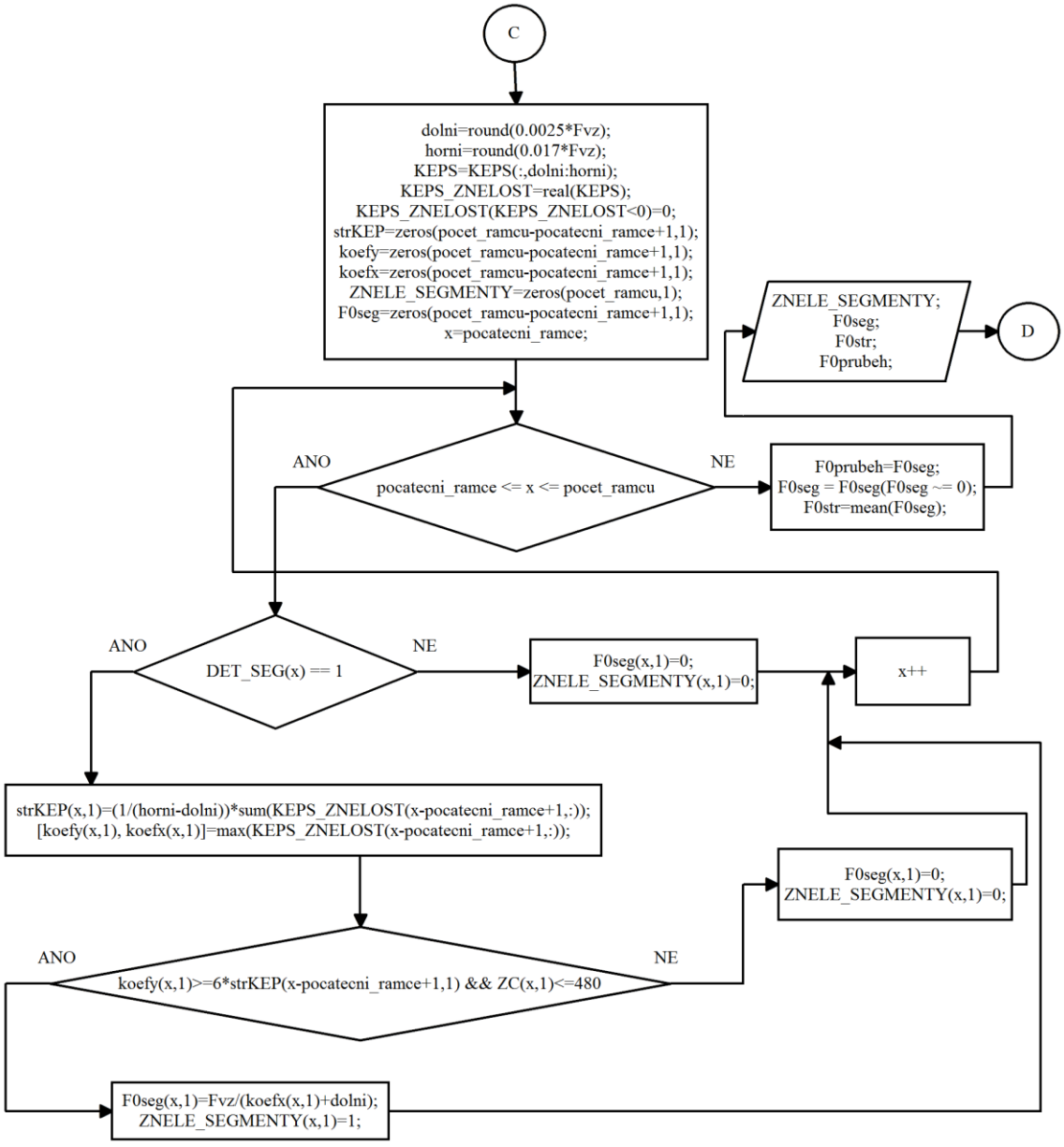
Sumační tvar diskrétní Fourierovy transformace (16) je nahrazen maticovým součinem transponované matice sloupců vzorků segmentů *SEGMENTY* a transponovaného řádkového vektoru komplexní exponenciály *EXPO_T* , cyklicky, pro nultý až *pocet_vz_1segmentu* – 1 koeficient *k*. Výsledkem je matice řádkových vektorů komplexních čísel *DFT*. Po ukončení cyklu je matice *DFT* přepočtena na novou matici logaritmovaného výkonového spektra *logPS*. V dalším kroku je proveden cyklus výpočtu inverzní Fourierovy transformace, pro nultý až *pocet_vz_1segmentu* – 1 koeficient *n*, přičemž matice *logPS* je v každém cyklu násobena transponovaným vektorem komplexní exponenciály *IDFT_EXP*. Výsledkem je matice řádkových vektorů kepstrálních koeficientů. Princip implementace algoritmu výpočtu výkonového kepstra je znázorněn na obrázku (obrázek 25).



Obrázek 25: Vývojový diagram implementace algoritmu výpočtu kepstrálních koeficientů.

Ze získaných reálných kepstrálních koeficientů jsou k další analýze použity pouze kladné koeficienty kepstra. Záporné koeficienty jsou nahrazeny nulovými hodnotami a jsou definovány přibližné kepstrální meze *horni* a *dolni*, které určují interval vymezující oblast pro hledání indexu výrazného kepstrálního maxima, který po přepočtu na frekvenci reprezentuje hodnotu základního tónu řeči F_0 a je také reprezentací znělosti daného segmentu. Řádkové vektory proměnné *KEPS* jsou dodatečně redukovány na délku určenou výše zmíněnými mezemi a uloženy do proměnné *KEPS_ZNELOST*. Cyklus algoritmu pro výpočet F_0 a určení znělosti ve všech řečově aktivních segmentech (kdy aktuální hodnota v registru *DET_SEG* je rovna jedné) nejprve vypočte střední hodnotu koeficientů *strKEPS* a určí maximální koeficient o souřadnicích *koefx* a *koefy*. Je-li nalezené maximum alespoň šestkrát větší (určeno experimentálně), než střední hodnota a zároveň,

je li hodnota počtu průchodů nulou daného segmentu (předpoklad periodicity pro znělou řeč s formantovou strukturou) menší než 480, pak se jedná o znělou řeč. Do registru znělosti segmentů *ZNELE_SEGMENTY* je na aktuální pozici poté zapsána hodnota 1 a příslušná souřadnice *koefx* je přepočtena na F_0 a uložena do proměnné *F0seg*. Pro segmenty, které byly klasifikovány jako řečově pasivní nebo neznělé je do sloupcových vektorů *ZNELE_SEGMENTY* a *F0seg* uložena hodnota 0. Průběh algoritmu je znázorněn na obrázku (obrázek 26).

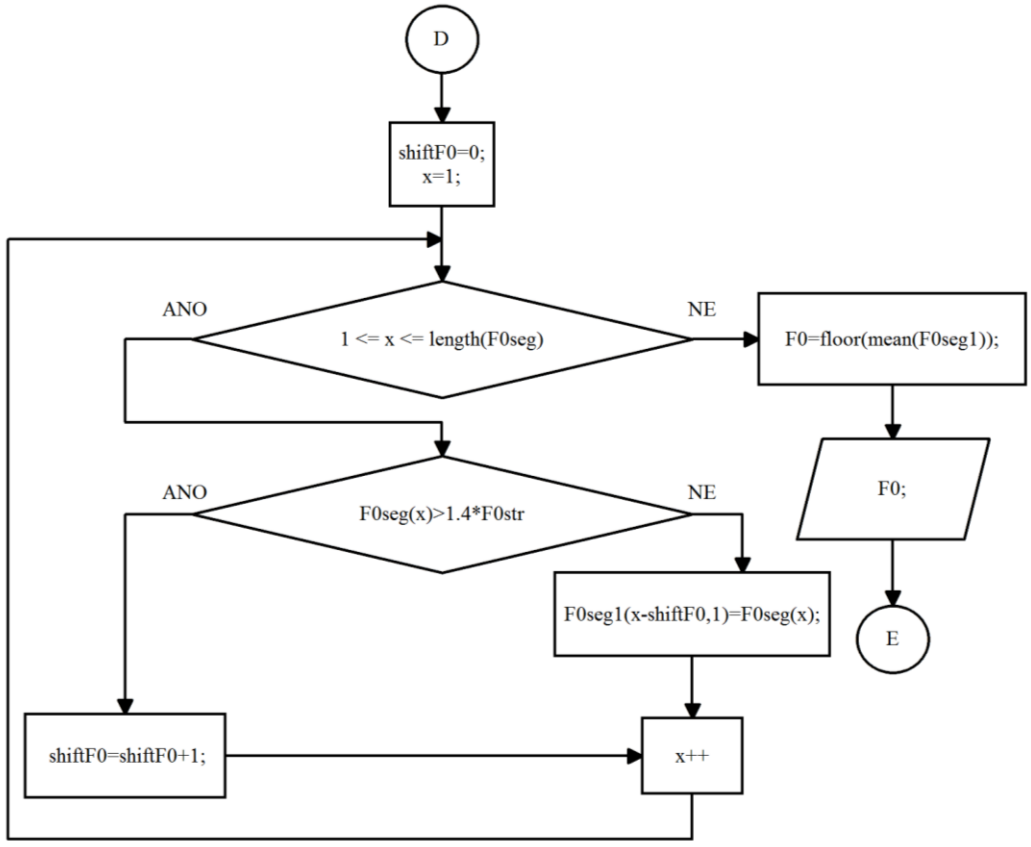


Obrázek 26: Vývojový diagram implementace algoritmu pro určení F_0 a znělosti segmentů.

Výstupem algoritmu pro výpočet F_0 a určení znělosti segmentu je registr znělých segmentů *ZNELE_SEGMENTY* , registr hodnot základní frekvence hlasu *F0seg* a jeho kopie *F0prubeh*, sloužící ke grafickému znázornění průběhu hodnot F_0 v segmentech. Nulové hodnoty z registru

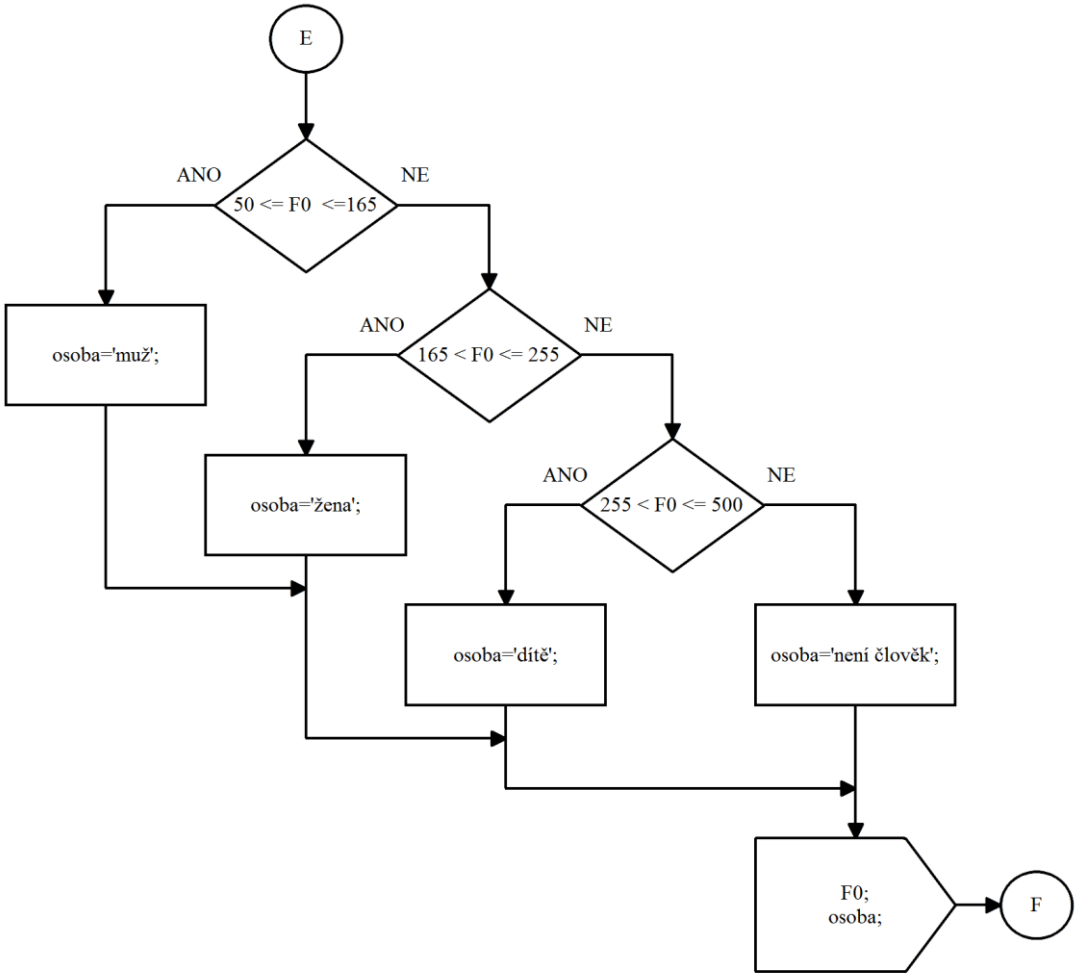
F0seg jsou po zápisu do vektoru *F0prubeh* odstraněny a je spočtena střední hodnota *F0str*. Zejména u segmentů náležících začátku a konci znělé promluvy bývají krátkodobě přítomny výrazně zvýšené hodnoty F_0, které by mohly významně ovlivnit výpočet odhadu průměrné F_0. Předpokladem je zde relativně neměnná hodnota základní frekvence řeči běžné promluvy, nikoliv zpěvu apod., kde mohou být velmi dynamické výkyvy F_0.

Získané hodnoty F_0 reprezentovány vektorem *F0seg* a z těchto získaná střední hodnota *F0str* jsou užity pro eliminaci výše zmíněných frekvenčních špiček (obrázek 27). Je zavedena nová proměnná *F0seg1*, do které jsou cyklicky zapisovány hodnoty F_0 jen v případě že daná hodnota základní frekvence nepřekračuje střední hodnotu *F0str* o 40 procent. Není li tato podmínka splněna, je inkrementována pomocná proměnna *shiftF0*, zajišťující postupné plnění sloupcového vektoru *F0seg1* hodnotami F_0. V posledním kroku jsou tyto hodnoty zprůměrovány a výsledné číslo je uloženo do proměnné *F0*.



Obrázek 27: Vývojový diagram implementace eliminování špičkových hodnot F_0.

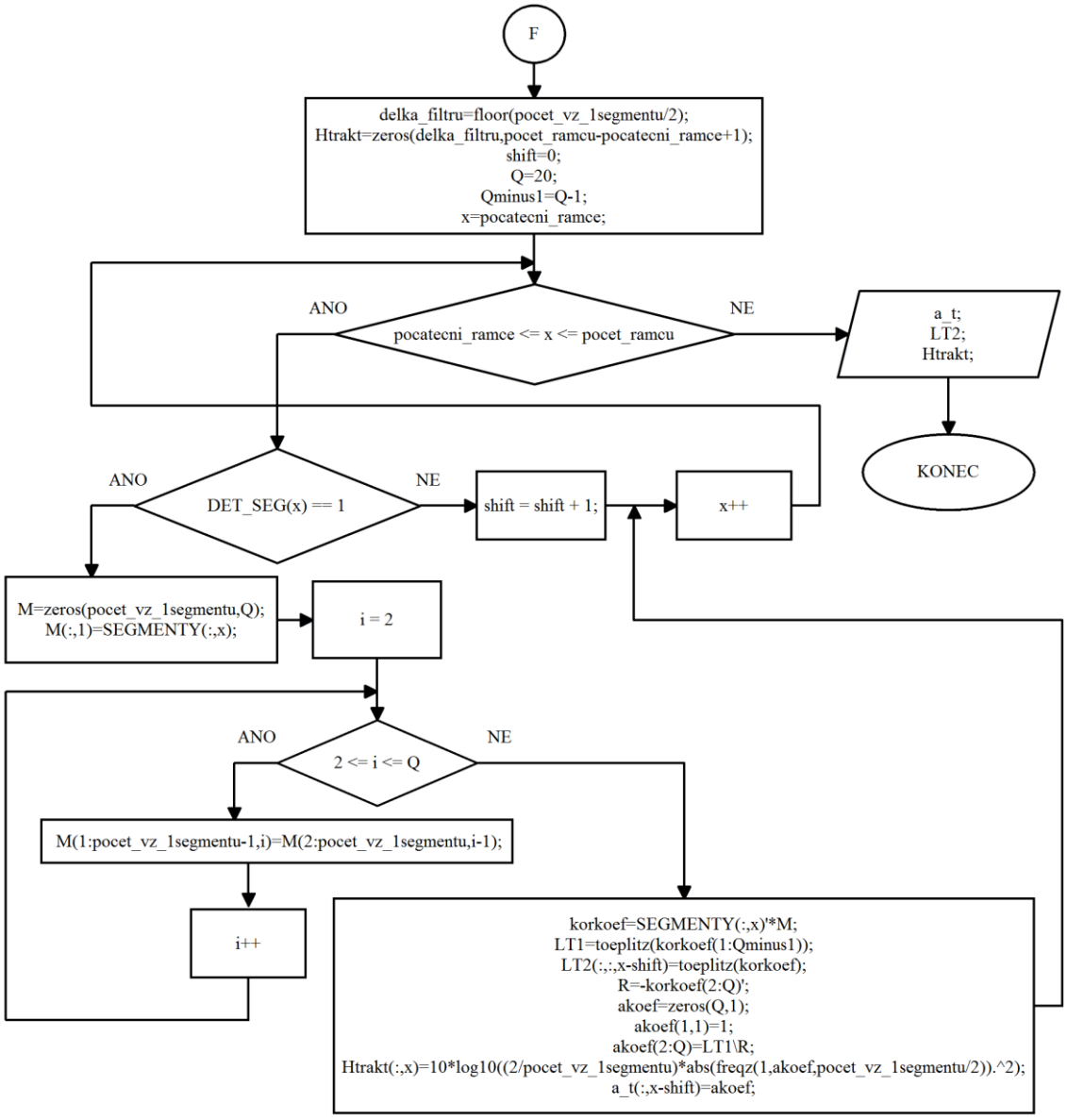
Po zjištění průměrné hodnoty základní frekvence hlasivek lze na jejím základě určit typ řečníka, a sice muže, ženu, nebo dítě. Klasifikace typu řečníka (obrázek 28) je provedena jednoduchým způsobem, na základě podmínek pro intervaly hodnot F_0. Spadá li hodnota proměnné *F0* do některého ze třech definovaných intervalů pro typ osoby, je do proměnné *osoba* tento typ zapsán. V opačném případě je do proměnné osoba zapsán text „není člověk“.



Obrázek 28: Vývojový diagram implementace algoritmu pro rozhodování o typu osoby.

Poslední částí funkce *detekce_a_mluvci* je výpočet *LPC* koeficientů za účelem dosazení do přenosové funkce charakteristiky řečového traktu dle vztahu (21), k získání vyhlazeného spektra a jsou také použity coby příznaky pro rozpoznání izolovaných slov pomocí metody *DTW*. Koeficienty lineární predikce jsou počítány pouze pro řečově aktivní segmenty a pro použití v oblasti rozpoznání izolovaných slov je zásadní kvalita provedené detekce. Problém nastává zejména u slov, majících uprostřed nějakou okluzívu, kdy je mezi znělou hláskou a okluzívou přítomna krátká pauza okolo deseti až dvaceti milisekund, což často vede k pořízení sekvence příznaků s vynecháním příznaků náležících právě této chybně určené pauze. Je určen řád predikce Q, roven 20. Cyklus pro výpočet predikčních koeficientů (obrázek 29) je řešen následujícím způsobem. Pro každý řečově aktivní segment je naplněna matice M o Q sloupcích, kde jednotlivé sloupce jsou reprezentovány signálem daného segmentu, posunutým o 0 až Q vzorků. Autokorelační koeficienty jsou vypočteny součinem transponovaného sloupce aktuálního segmentu z matice *SEGMENTY* s maticí M. Výsledkem je řádkový vektor autokorelačních koeficientů *korkoef*. Z nultého až $Q - 1$ korelačního koeficientu je následně vytvořena Töplitzova matice *LT1* pomocí příkazu *toeplitz* a obdobně pak, ze všech Q autokorelačních koeficientů aktivních segmentů

třírozměrná matice *LT2*, která je cyklicky naplňována a je využita při rozpoznávání izolovaných slov metodou *DTW*, v rámci výpočtu Itakurovy míry vzdálenosti. Predikční koeficienty každého aktivního segmentu jsou vypočteny řešením soustavy rovnic dle vztahu (23), součinem inverze matice *LT1* se se záporně vzatým transponovaným řádkovým vektorem *korkoef* , redukovaným na první až *Q* - tý autokorelační koeficient. Výsledkem je sloupcový vektor *LPC* koeficientů *akoef* o *Q* složkách, který je v každém cyklu ukládán do proměnné *a_t*, čímž vzniká časová sekvence příznaků aktivních segmentů. Dosazením vektoru *akoef* do přenosové charakteristiky řečového traktu *Htrakt* je získán vyhlazený odhad výkonového spektra segmentu v decibelech.



Obrázek 29: Vývojový diagram implementace výpočtu *LPC* koeficientů.

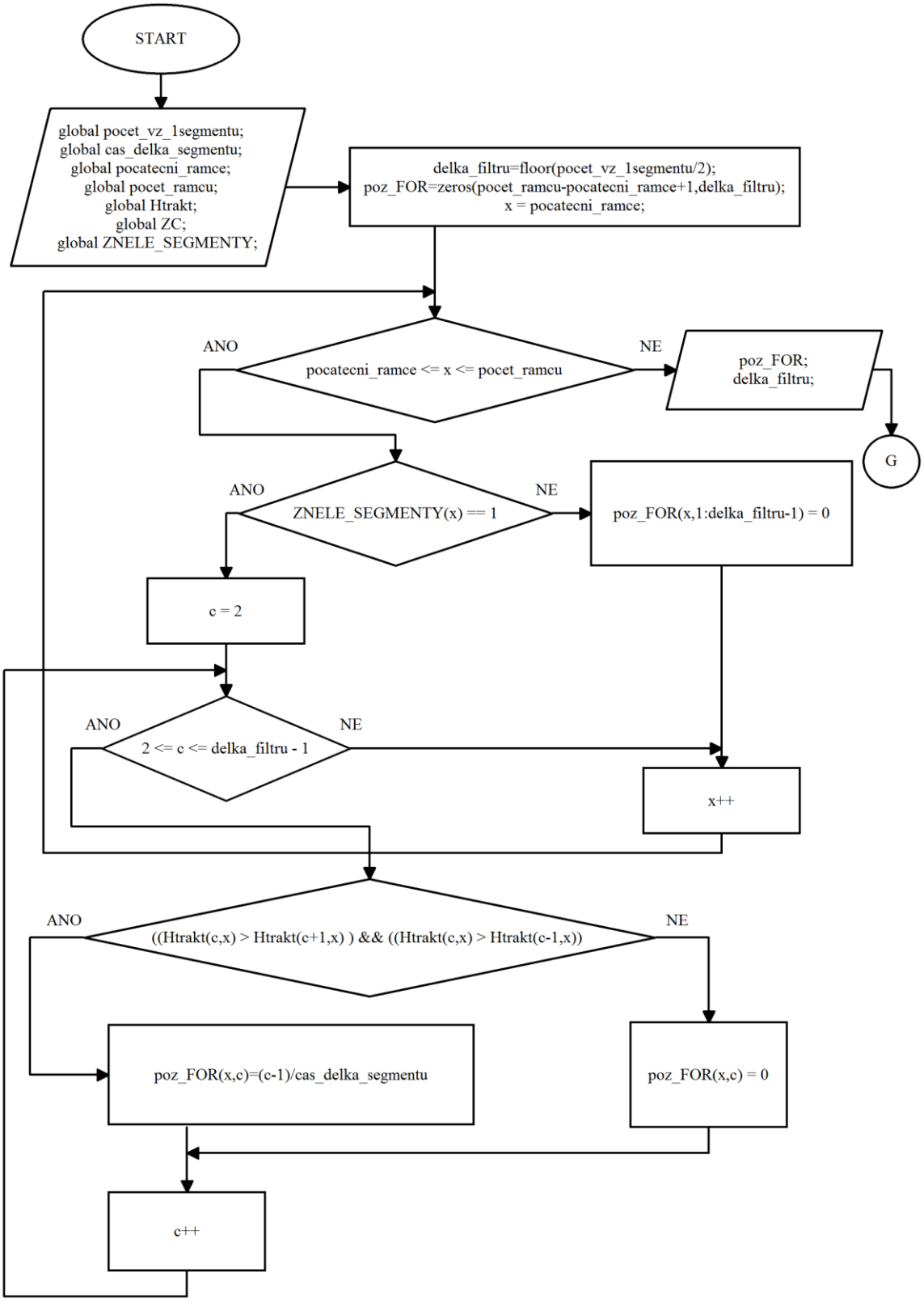
7.2.3 Odhad samohlásek *a*, *e*, *i*, *o*, *u*

Odhad samohlásek *a*, *e*, *i*, *o*, *u* je proveden prostřednictvím funkce *odhad_hlasek*, volané tlačítkem *zobrazit řetězec znaků* v panelu *Výsledky rozpoznání řeči*. Předpokladem pro spuštění funkce je provedení detekce řeči (pomocná proměnná *detekovano* je rovna 1) a tím i získání vyhlazených spektrálních charakteristik řečového traktu jednotlivých segmentů v matici *Htrakt*. Vstupní a výstupní proměnné funkce jsou uvedeny v tabulce (tabulka 5).

Tabulka 5: Vstupní a výstupní proměnné funkce *odhad_hlasek*.

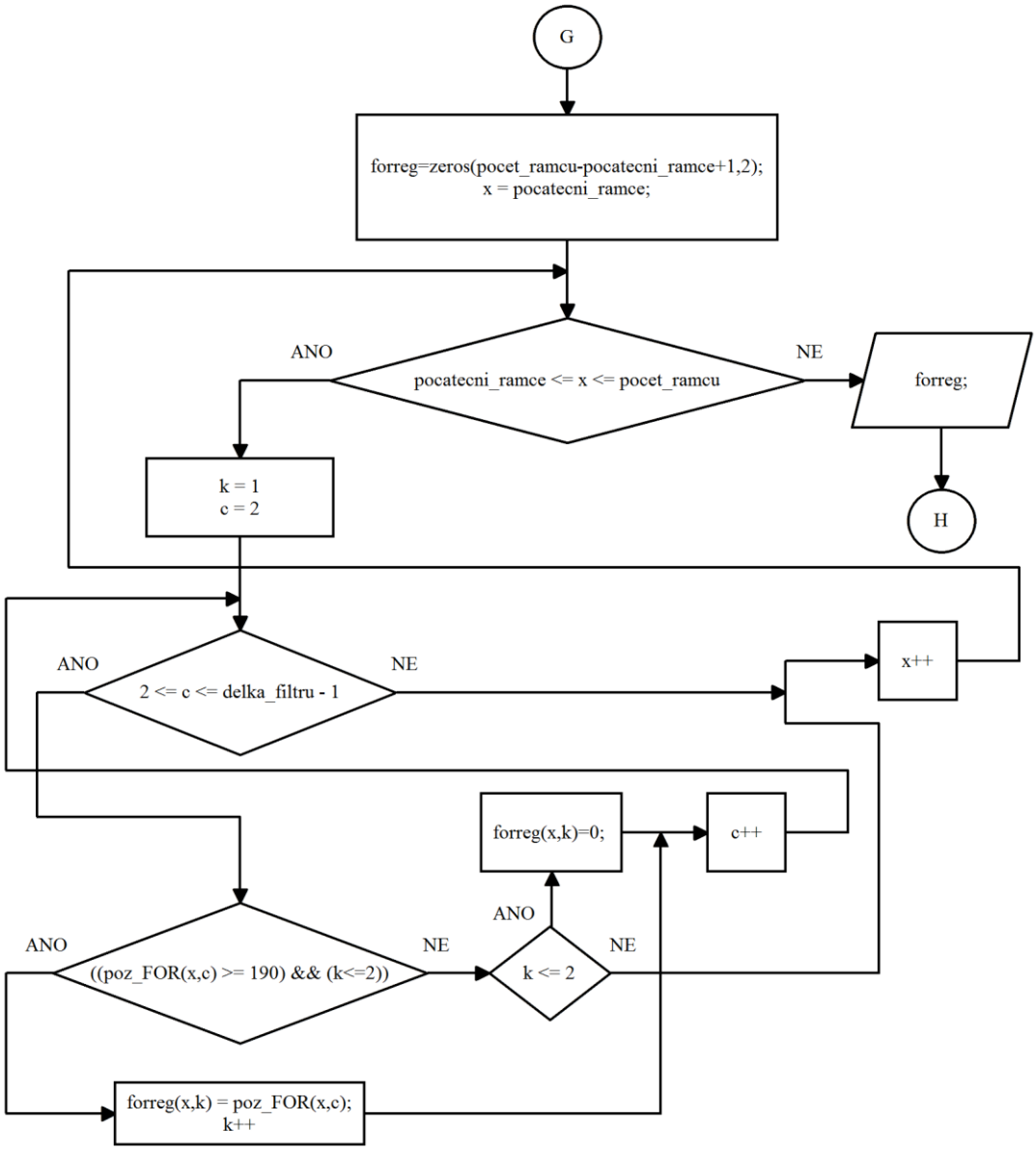
Vstupní proměnná	Výstupní proměnná
global *pocet_vz_1segmentu*	global *hlaska*
global *cas_delka_segmentu*	
global *pocatecni_ramce*	
global *pocet_ramcu*	
global *Htrakt*	
global *ZC*	
global *ZNELE_SEGMENTY*	

Prvním krokem je nalezení indexů lokálních maxim (obrázek 30), tedy formantů, v jednotlivých sloupcích hodnot výkonových spekter segmentů *Htrakt*. Cyklus hledá pro všechny znělé segmenty (aktuální hodnota v registru *ZNELE_SEGMENTY* musí být rovna jedné) lokální maxima v aktuálním sloupci matice *Htrakt*, ve všech řádcích, jejichž počet je dán proměnnou *delka_filtru*. Je li nalezeno lokální maximum, pak je jeho index řádku přepočten na frekvenci a uložen do proměnné *poz_FOR* na pozici sloupce o hodnotě příslušného indexu, řádek pak odpovídá pořadí analyzovaného segmentu. V případě nepřítomnosti maxima je tímtéž způsobem na dané aktuální pozici zapsána hodnota 0. Je li segment neznělý, pak jsou všechny sloupce v daném řádku proměnné *poz_FOR* vyplněny nulami.



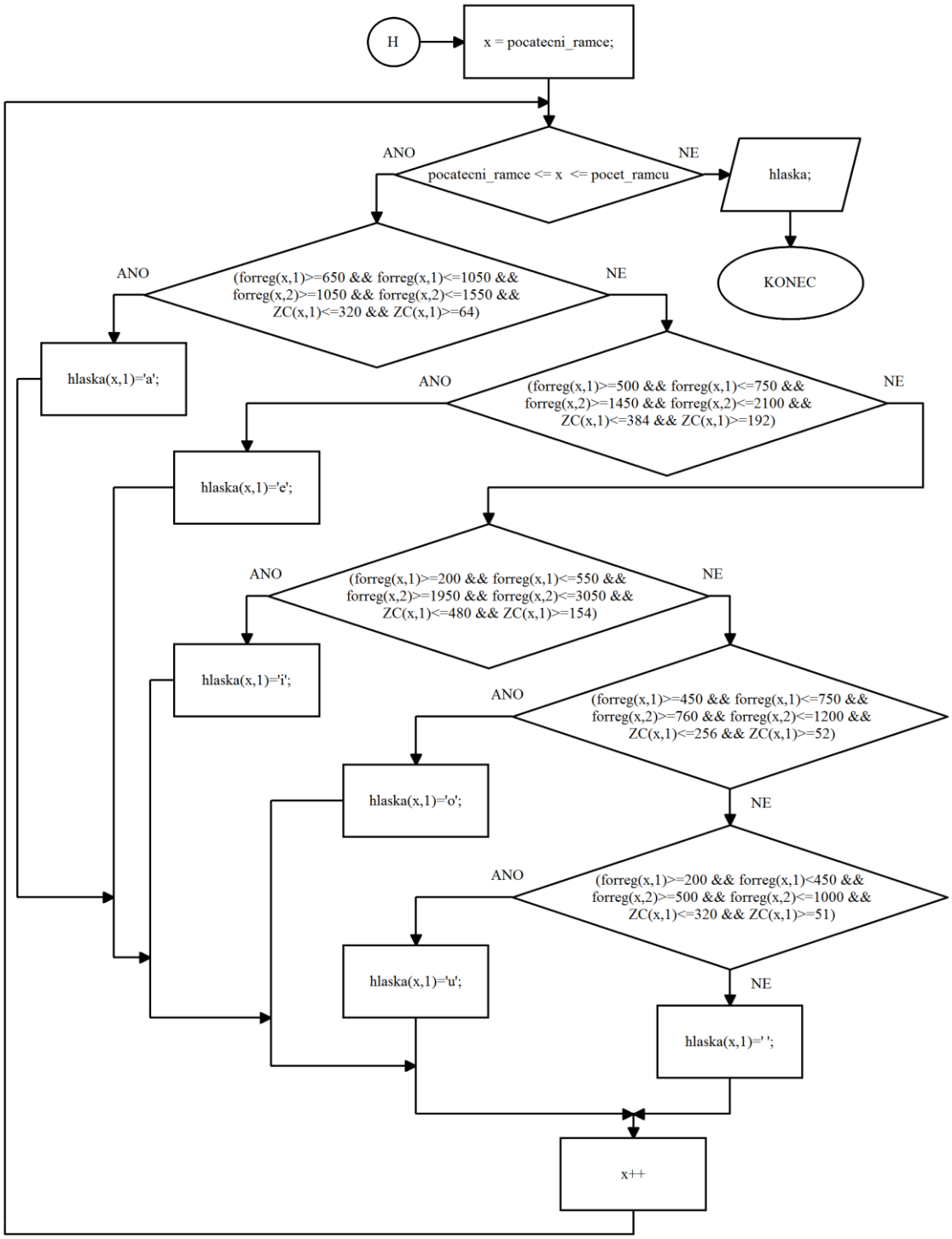
Obrázek 30: Vývojový diagram implementace algoritmu pro hledání formantů.

Dalším krokem je následná extrakce formantů F_1 a F_2 (obrázek 31) ze všech získaných extrémů, jenž jsou uloženy v proměnné *poz_FOR*. Je definována nová matice, proměnná *forreg* o dvou sloupcích, jejíž počet řádků je roven počtu analyzovaných segmentů. Cyklus pro hledání prvních dvou formantů hledá v jednotlivých řádcích a všech sloupcích matice *poz_FOR* dvě první maximální hodnoty frekvence, větší než 190 Hz. Pomocná proměnná *k* o počáteční hodnotě 1 udává index sloupce proměnné *forreg*, do kterého jsou uloženy hodnoty F_1 a F_2 a je inkrementována o hodnotu 1 ihned po uložení hodnoty formantu do *k* - tého sloupce. Je li hodnota proměnné k větší než 2, tím pádem jsou li nalezeny právě dva první formanty, je ukončen cyklus pro hledání v aktuálním řádku matice *poz_FOR* a tímtéž způsobem jsou analyzovány řádky následující. V případě neznělých segmentů jsou do obou sloupců v daném řádku proměnné *forreg* zapsány hodnoty 0. Výstupem je proměnná *forreg*, naplněná hodnotami F_1 a F_2.



Obrázek 31: Vývojový diagram implementace algoritmu pro extrakci formantů F_1 a F_2.

Poslední částí funkce *odhad_hlasek* je heuristický odhad znělých samohlásek (obrázek 32) na základě formantů *F*1, *F*2 získaných z proměnné *forreg* a počtu průchodů nulou z proměnné *ZC*. Vyhovují li tyto veličiny svými hodnotami některé z definovaných podmínek, pak je jim přiřazen význam konkrétní hlásky a tato je následně ukládána pod příslušným znakem do řádků proměnné *hlaska*. V případě nepřiřazení hlásky danému segmentu je na danou pozici zapsán prázdný znak.



Obrázek 32: Vývojový diagram implementace algoritmu pro odhad hlásek *a*,*e*,*i*,*o*,*u*.

7.2.4 Rozpoznávání izolovaných slov

Rozpoznávání izolovaného slova je založeno na metodě *DTW*. Jsou li k dispozici referenční sekvence *LPC* příznaků *a_tR* pro jednotlivá referenční slova a byla li provedena detekce testované promluvy, pak lze prostřednictvím funkce *vypocti_DTW* provést klasifikaci řečeného slova. Vstupní a výstupní proměnné funkce jsou uvedeny v tabulce (tabulka 6).

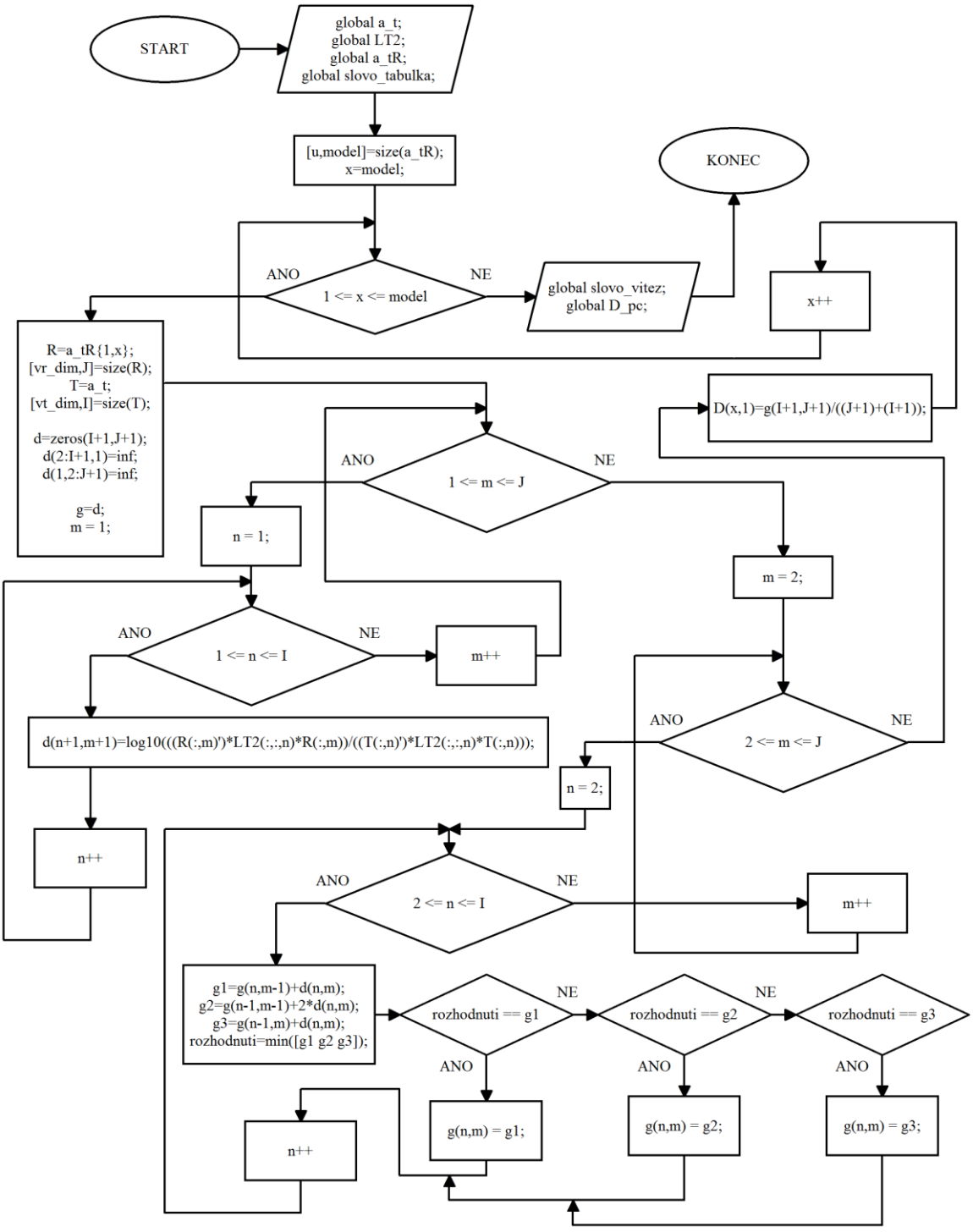
Tabulka 6: Vstupní a výstupní proměnné funkce *vypocti_DTW*.

Vstupní proměnná	Výstupní proměnná
global *a_t*	global *slovo_vitez*
global *LT2*	global *D_pc*
global *a_tR*	
global *slovo_tabulka*	

Nyní bude popsán princip algoritmu funkce *vypocti_DTW* (obrázek 33). Nejprve je načtena matice sekvence vektorů *LPC* příznaků testovaného slova *a_t*, trojrozměrná matice *LT2* reprezentující jednotlivé Töplitzovy matice z *Q* autokorelačních koeficientů náležících segmentům testovaného slova, pole sekvencí příznaků jednotlivých referenčních slov *a_tR* a pole názvů všech zadaných referenčních slov *slovo_tabulka*. Poté je zjištěn počet buněk (sloupců) pole proměnné *a_tR* za účelem získání počtu všech referenčních *LPC* obrazů, které budou následně porovnávány s testovaným obrazem *a_t*. Počet těchto obrazů je zapsán do proměnné *model*. V každém cyklu výpočtu vzdálenosti je z pole *a_tR* vyňata aktuální buňka sekvence příznaků a uložena do matice *R* o *Q* řádcích a *J* sloupcích, testovaná sekvence příznaků je uložena do matice *T* o *Q* řádcích a *I* sloupcích. Poté jsou mezi *m*-tým až *J*-tým vektorem aktuální referenční sekvence a mezi *n* – tým až *I* tým vektorem sekvence testované, vypočteny vzájemné lokální vzdálenosti *d*, dle Itakurovy míry. Dalším krokem je výpočet akumulovaných vzdáleností *g* dle vztahu pro lokální omezení typu I dle rovnice (50), jimiž je naplněna stejnojmenná matice. Celkové vzdálenost mezi testovanou sekvencí příznaků a sekvencemi příznaků vzorových slov jsou ukládány do sloupcového vektoru *D*. Po ukončení algoritmu *DTW* je počet řádků proměnné *D* roven hodnotě proměnné *model*. Následně je ze všech celkových vzdáleností určeno minimum a jeho souřadnice, které rovněž slouží k přiřazení vítězného slova zapsaného na dané pozici v proměnné *slovo_tabulka*. Minimální vzdálenost je přepočtena na procentuální podobnost *D_pc* a spolu s vítězným slovem jsou tyto údaje vypsány do příslušných textových polí v panelu *Převod řeči na text*. Práh pro určení vítězného slova je určen na alespoň 25 procent celkové podobnosti.

Implementace části kódu klasifikace slova funkce *vypocti_DTW*

```
[Mhodnota,Mmodel]=min(D);
if min(D)<=0.75
slovo_vitez=slovo_tabulka{Mmodel,1};% slovo s nejmensi vzdalenosti vuci vsem referencim
D_pc = floor(100*(1-min(D)));% prepocet vitezne vzdalenosti na procenta
set(handles.podobnost,'String',D_pc);% vypis
set(handles.vypis_slovo_DTW,'String',slovo_vitez);
else
slovo_vitez='nerozpoznáno';
set(handles.vypis_slovo_DTW,'String',slovo_vitez);
set(handles.podobnost,'String',[]);
end
```



Obrázek 33: Vývojový diagram implementace funkce *vypocti_DTW*.

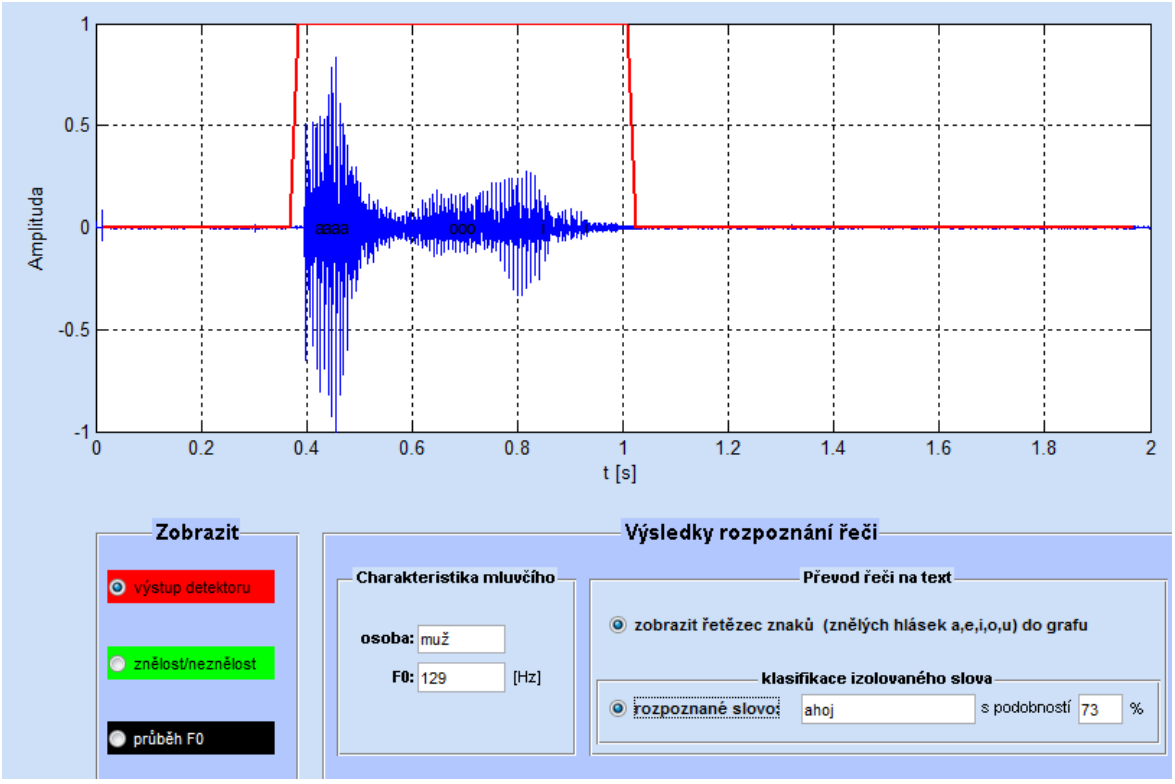
7.3 Analýza výsledků

7.3.1 Účinnost intenzitního detektoru

Objektivní metodou pro vyhodnocení účinnosti detekce může být např. výpočet poměru signálu vůči šumu v decibelech (také zkr. *SNR*, z angl signal to noise ratio), daný vztahem

$$SNR = 10 \log \left\{ \frac{\sum_{n=0}^{N-1} s^2(n)}{\sum_{n=0}^{N-1} [s(n) - \hat{s}(n)]^2} \right\}, \tag{60}$$

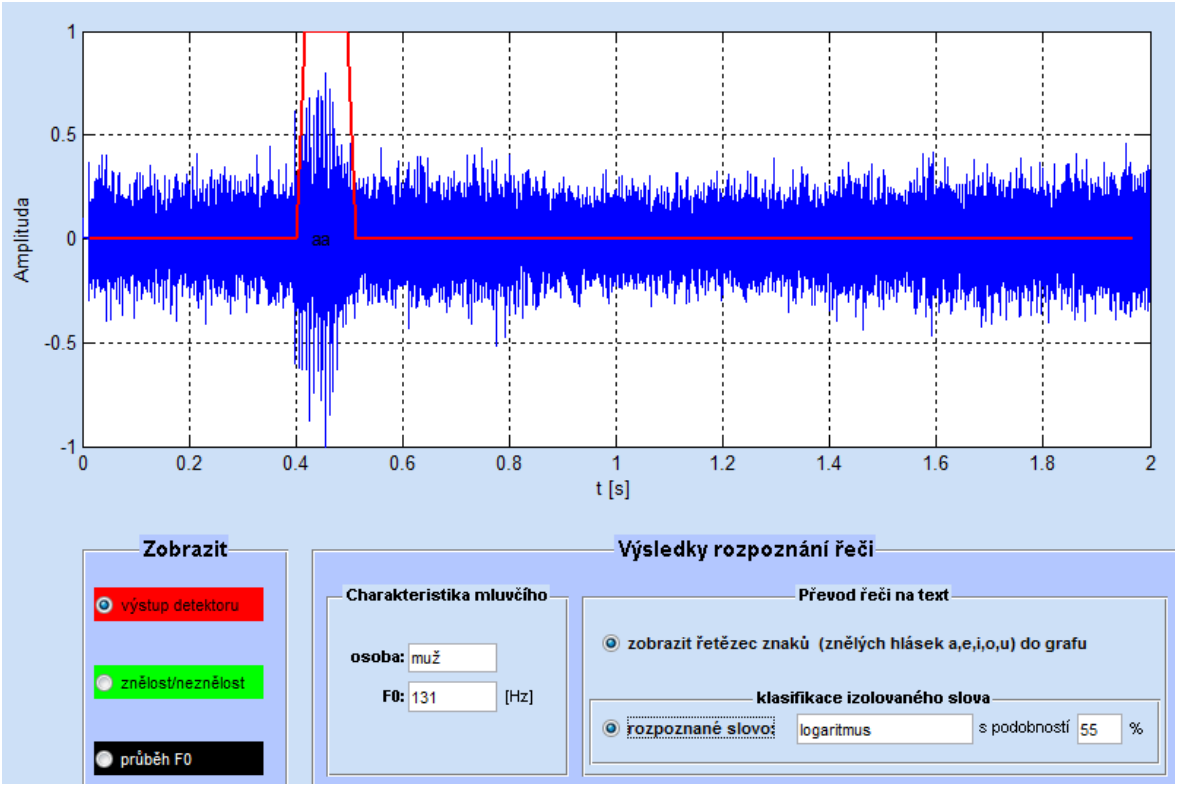
kde $s(n)$ je čistý řečový signál a $\hat{s}(n)$ je řečový signál ovlivněný hlukem. Detektory založené na měření změn intenzity či energie jsou obecně citlivé na dynamické změny rušivého prostředí, proto jsou vhodné tam, kde je nežádoucí hluk spíše stacionární povahy. Na obrázku (obrázek 34) je znázorněn průběh detekce čistého signálu slova „ahoj“, s výpisem odhadu znělých hlásek, klasifikace izolovaného slova (tatáž osoba pro vzorové i testované promluvy), určení typu osoby (muž) a hodnota F_0. Všechny zjištěné údaje korespondují s danými předpoklady.



Obrázek 34: Detekce čistého řečového signálu slova „ahoj".

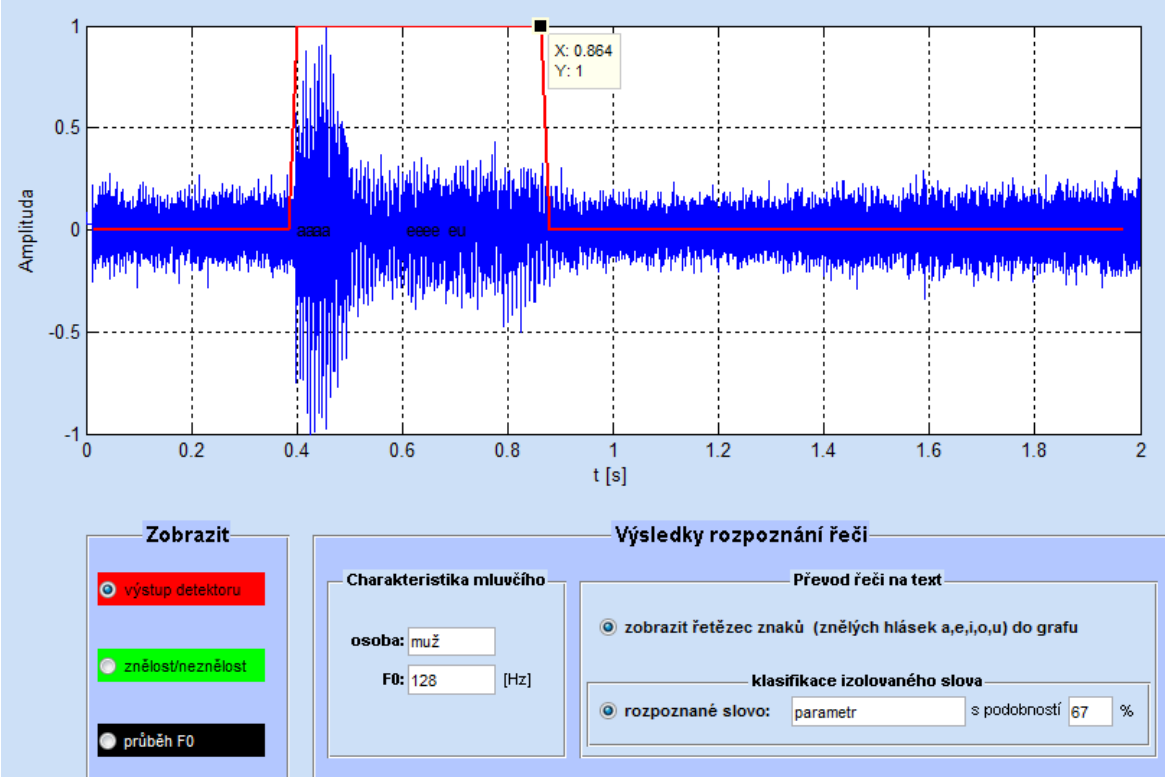
V případě signálu, jehož celkový poměr *SNR* je okolo 4 dB a méně, může u některých částí promluvy s nižší energií dojít k chybě detekce v okamžiku, kdy je řečový signál aktivních segmentů překryt stacionárním hlukem. Při použití energetického detektoru by bylo možno detekovat velmi malé změny v intenzitě signálu, avšak za cenu mylné detekce pasivních segmentů a jejich následného prohlášení za řečově aktivní. Na obrázku (obrázek 35) je znázorněna tatáž

promluva jako v předchozím případě, tentokrát je ovšem na řečový signál superponován hluk fénu. Poměr *SNR* je roven 1 dB a kromě samohlásky „a“ lze pozorovat neúspěšnou detekci zbytku slova, v porovnání s detekcí čistého signálu. Odhad typu osoby (muž) koresponduje se skutečností, kepstrální metoda výpočtu F_0 se zdá být účinná i za přítomnosti hluku. Rozpoznání izolovaného slova je na základě detekce pouhé části promluvy zcela chybné.



Obrázek 35: Detekce řečového signálu slova „ahoj", s odstupem *SNR* o hodnotě 1 dB.

Při poměru *SNR* o hodnotě 5 dB (obrázek 36) je kvalita detekce promluvy slova „ahoj“ téměř srovnatelná s kvalitou detekce čistého signálu. Nedostatkem je nepřesná detekce konce slova v čase asi 0,86 s, přičemž skutečný čas ukončení promluvy lze v čistém signálu pozorovat kolem jedné sekundy (obrázek 34). Odhad typu osoby a výpočet hodnoty F_0 korespondují s analýzou čistého signálu, výpis odhadu znělých hlásek je správný pouze pro část promluvy reprezentovanou hláskou „a“. Klasifikace izolovaného slova je neúspěšná z důvodu ovlivnění signálu hlukem z fénu.



Obrázek 36: Detekce řečového signálu slova „ahoj“, s odstupem SNR o hodnotě 5 dB.

Z výše uvedených závěrů plyne, že detekce řeči v rušném prostředí je poměrně problematickou záležitostí, má li být založena na veličinách jako intenzita či energie, které jsou počítány v časové oblasti. V případě, že by rušivý signál disponoval velkými výkyvy intenzity, docházelo by k naprosto mylné detekci úseků, kde by řeč nemusela být přítomna. Z tohoto důvody byly v rámci analýzy řečového signálu pořízeny nahrávky s velkým poměrem *SNR*.

7.3.2 Úspěšnost odhadu samohlásek *a*, *e*, *i*, *o*, *u*

Úspěšnost odhadu samohlásek *a*, *e*, *i*, *o*, *u*, je určena poměrem správných znaků z výpisu textového řetězce, ke všem vypsaným znakům, náležícím danému detekovanému úseku samohlásky (ukázka v příloze II). Je li tento poměr menší nebo roven 0,5, pak je odhad neúspěšný. K dispozici jsou nahrávky deseti mluvčích (pět mužů a pět žen), kde každému mluvčímu náleží deset měření promluvy samohlásek „a, e, i, o, u“. Mluvčí jsou značeni $m1$ až $m10$, přičemž $m1$ až $m5$ jsou muži a $m6$ až $m10$ ženy. Pro účely statistického zpracování záznamů je vytvořena tabulka absolutních četností správných odhadů jednotlivých samohlásek (tabulka 7). Poslední řádek a poslední sloupec tabulky jsou vyplněny hodnotami relativních četností, vyjadřujících celkové procentuální úspěšnosti jednotlivých odhadů. Celková procentuální úspěšnost odhadu konkrétní samohlásky řečené všemi mluvčími $úsp(sam)$ je vypočtena dle vztahu

$$\text{úsp}(sam)\ [\%] = \frac{suma\ všech\ správných\ odhadů\ konkrétní\ samohlásky}{počet\ měření\ samohlásky\ u\ všech\ osob\ (100\ měř.)} \cdot 100\,, \qquad (61)$$

obdobně pak celková procentuální úspěšnost odhadu všech samohlásek řečených jedním mluvčím *úsp*(*m*) je vypočtena dle vztahu

$$\text{úsp}(m)\ [\%] = \frac{suma\ správných\ odhadů\ všech\ samohlásek\ (mluvčí\ m)}{počet\ měření\ všech\ samohlásek\ (jeden\ ml.,\ 50\ měř.)} \cdot 100\,. \qquad (62)$$

Tabulka 7: Výsledky měření úspěšnosti odhadu samohlásek *a*, *e*, *i*, *o*, *u*.

samohláska	*m*1	*m*2	*m*3	*m*4	*m*5	*m*6	*m*7	*m*8	*m*9	*m*10	*úsp*(*sam*)	*úsp*(*sam*) [%]
a	0	0	8	10	10	9	10	5	5	9	66	66
e	3	10	10	10	10	7	9	10	6	10	85	85
i	9	9	10	10	10	10	10	10	10	10	98	98
o	10	7	7	4	10	10	9	1	7	8	73	73
u	10	10	10	4	2	8	6	9	1	7	67	67
úsp(*m*)	32	36	45	38	42	44	44	35	29	44		
úsp(*m*) [%]	64	72	90	76	84	88	88	70	58	88		

Z výsledků (tabulka 7) lze vyčíst mezi některými z mluvčích *m* (modře muži, červeně ženy) znatelně rozdílné hodnoty celkové procentuální úspěšnosti. To je dáno zejména experimentálním nastavením mezí intervalů měřených veličin (F_1, F_2 a počet průchodů nulou), které prostřednictvím podmínek přiřazují segmentům řeči význam dané samohlásky. S tímto souvisí také jistá variabilita v rozměrech řečového ústrojí a aktuální pozice řečového traktu každého z mluvčích, neboť právě tyto vlastnosti mají vliv na polohu formantů ve spektru řeči. V rámci této analýzy se nepředpokládá závislost úspěšnosti odhadu samohlásky na pohlaví. Nejmenší celkovou úspěšností odhadu disponují hlásky „a“ a „u“, které jsou z celkových 100 měření odhadnuty 66 krát, tedy v 66 procentech, resp. 67 krát, tedy v 67 procentech počtu měření. V případě mluvčích *m*1 a *m*2 (muži) je hláska „a“ v rámci všech deseti měření v drtivé většině segmentů chybně prohlášena za hlásku „o“. Samohláska „u“ je zejména v případech promluvy mluvčím *m*9 (žena, pouze 1 správný odhad z 10) a mluvčím *m*5 (muž, pouze 2 správné odhady) chybně prohlášena za hlásku „i“. Úspěšnost odhadu samohlásky „o“ (73 procent) je taktéž ovlivněna ostrostí nastavených podmínek a zejména v případě mluvčího *m*8 (žena, pouze 1 správný odhad) je chybně prohlášena za samohlásku „a“. Nejlepší celkové výsledky odhadu jsou pořízeny u samohlásek „e“ (úspěšnost odhadu 85 procent) a „i“ (98 procent), přičemž samohláska „e“ je pouze v případě mluvčího *m*1 (muž) úspěšně odhadnuta jen 3 krát z deseti měření a u zbylých mluvčích se pohybují jednotlivé absolutní četnosti mezi šesti až deseti úspěšnými odhady. Četnosti výskytu správných odhadů samohlásky „i“ nabývají u jednotlivých mluvčích hodnot 9 až 10, jedná se tedy o nejvíce úspěšný odhad v rámci všech testovaných samohlásek. Rozsah celkové úspěšnosti odhadu všech samohlásek konkrétním mluvčím se pohybuje od 58 procent do 90 procent.

7.3.3 Úspěšnost rozpoznání izolovaných slov

Pro rozpoznání izolovaně řečených slov na základě metody *DTW* jsou k dispozici vzorové nahrávky slov *ahoj*, *olej*, *sinus*, *parametr*, *logaritmus* a *humanismus*, řečené jedním mužským hlasem. Od všech testovaných mluvčích $m1$ až $m10$ je pořízeno deset nahrávek každého slova. Výsledky úspěšnosti rozpoznání slov jsou pro každé pohlaví vypočteny zvlášť (tabulka 8 a tabulka 9). Neodpovídá li výsledek rozpoznání analyzovanému slovu, popř. je li výstupem rozpoznání hláška *nerozpoznáno*, pak se jedná o neúspěšný výsledek měření. Celková procentuální úspěšnost rozpoznání konkrétního slova řečeného všemi mluvčími stejného pohlaví je dána vztahem

$$\text{úsp}(sl)\ [\%] = \frac{počet\ všech\ rozpoznaných\ nahrávek\ slova}{50} \cdot 100 \qquad (63)$$

a celková procentuální úspěšnost rozpoznání všech slov konkrétním mluvčím m je vypočtena dle vztahu

$$\text{úsp}(m\_sl)\ [\%] = \frac{počet\ všech\ rozpoznaných\ nahrávek\ řečených\ osobou\ m}{60} \cdot 100 \qquad (64)$$

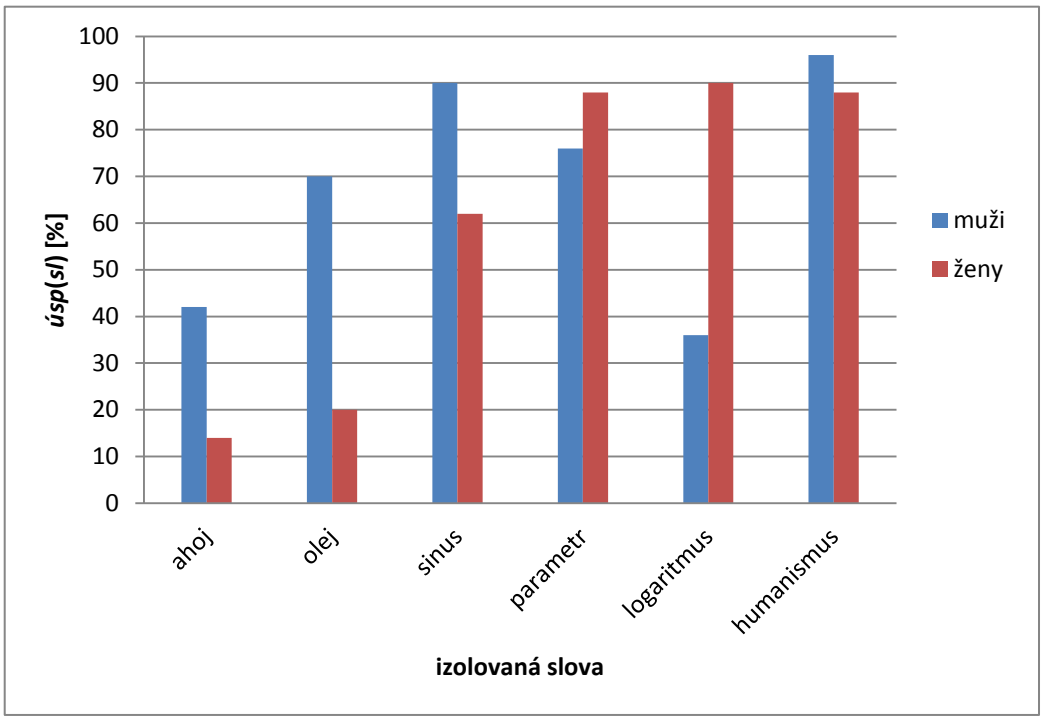
Tabulka 8: Výsledky měření úspěšnosti rozpoznání izolovaných slov - muži.

slovo	$m1$	$m2$	$m3$	$m4$	$m5$	*úsp*(*sl*)	*úsp*(*sl*) [%]
ahoj	0	0	8	8	5	21	42
olej	10	1	10	5	9	35	70
sinus	10	5	10	10	10	45	90
parametr	0	9	10	10	9	38	76
logaritmus	5	2	2	7	2	18	36
humanismus	9	9	10	10	10	48	96
úsp(*m_sl*)	34	26	50	50	45		
úsp(*m_sl*) [%]	57	43	83	83	75		

Tabulka 9: Výsledky měření úspěšnosti rozpoznání izolovaných slov - ženy.

slovo	$m6$	$m7$	$m8$	$m9$	$m10$	*úsp*(*sl*)	*úsp*(*sl*) [%]
ahoj	1	6	0	0	0	7	14
olej	1	5	0	0	4	10	20
sinus	10	7	4	1	9	31	62
parametr	10	10	5	10	9	44	88
logaritmus	10	10	7	10	8	45	90
humanismus	9	8	8	9	10	44	88
úsp(*m_sl*)	41	46	24	30	40		
úsp(*m_sl*) [%]	68	77	40	50	67		

Častým důvodem chybného rozpoznání některých slov je zejména jejich podobnost, velkou roli zde zřejmě sehrává i intonace, barva hlasu a kvalita artikulace jednotlivých mluvčích. V případě mluvčích obou pohlaví dochází k časté chybné klasifikaci, tedy záměně slov „ahoj“ za „olej“ (u žen častěji), či „logaritmus“ za „humanismus“ (častěji u mužů). Nejmenší celkové procentuální úspěšnosti rozpoznání slova jsou u mužského pohlaví zastoupeny 36 procenty u slova „logaritmus“ a 42 procenty u slova „ahoj“, u ženského pak pouze 14 procenty u slova „ahoj“ a 20 procenty u slova „olej“. U mužské části mluvčích jsou nejlépe rozpoznatelná slova „olej“ (úspěšnost 70 %), „sinus“ (úspěšnost 90 %), parametr (úspěšnost 76 %) a humanismus (úspěšnost 96 %), u ženské části pak „sinus“ (úspěšnost 62 %), „parametr“ (úspěšnost 88 %), „logaritmus“ (úspěšnost 90 %) a „humanismus“ (úspěšnost 88 %). Úspěšnost rozpoznání všech slov řečených konkrétním mluvčím (zaokrouhleno na celé číslo) se pro mužskou část pohybuje od 43 % (mluvčí *m*2) do 83 % (mluvčí *m*3 a *m*4), pro ženskou část pak od 40 % (mluvčí *m*8) do 77 % (mluvčí *m*7). Pro přehlednost jsou výsledky úspěšnosti rozpoznání jednotlivých slov uvedeny v grafické podobě (obrázek 37).



Obrázek 37: Histogram procentuální úspěšnosti rozpoznání izolovaných slov pro obě pohlaví.

7.3.4 Úspěšnost rozpoznání typu mluvčího

Rozpoznání typu mluvčího je odvozeno z vyhodnocení průměrné hodnoty F_0. Mluvčímu je poté přidělen jeden ze tří typů osob, konkrétně *muž*, *žena*, nebo *dítě*. Měření jsou provedena na izolovaných slovech *ahoj*, *olej*, *sinus*, *parametr*, *logaritmus* a *humanismus*, přičemž počet měření je stejný jako v případě rozpoznávání izolovaných slov. Výsledky měření jsou rozděleny

zvlášť pro muže (tabulka 10) a ženy (tabulka 11). Celková procentuální úspěšnost rozpoznání typu mluvčího prostřednictvím promluvy konkrétního slova je vypočtena dle vztahu

$$\text{úsp}(typ)\ [\%] = \frac{počet\ všech\ úspěšných\ měření\ (konkrétní\ slovo)}{50} \cdot 100 \qquad (65)$$

a procentuální úspěšnost rozpoznání typu mluvčího u konkrétní osoby *m* je dána vztahem

$$\text{úsp}(m\_typ)\ [\%] = \frac{počet\ úspěšných\ měření\ (mluvčí\ m,\ \ všechna\ slova)}{60} \cdot 100\,. \qquad (66)$$

Tabulka 10: Výsledky měření úspěšnosti rozpoznání typu mluvčího – muži.

slovo	*m*1	*m*2	*m*3	*m*4	*m*5	*úsp*(*typ*)	*úsp*(*typ*) [%]
ahoj	10	10	10	10	7	47	94
olej	10	10	10	10	7	47	94
sinus	10	9	8	10	6	43	86
parametr	10	10	10	10	7	47	94
logaritmus	10	10	10	10	9	49	98
humanismus	10	10	10	10	9	49	98
úsp(*m_typ*)	60	59	58	60	45		
úsp(*m_typ*) [%]	100	98	97	100	75		

Tabulka 11: Výsledky měření úspěšnosti rozpoznání typu mluvčího – ženy.

slovo	*m*6	*m*7	*m*8	*m*9	*m*10	*úsp*(*typ*)	*úsp*(*typ*) [%]
ahoj	6	10	7	6	10	39	78
olej	5	10	6	9	10	40	80
sinus	2	8	5	6	10	31	62
parametr	5	10	5	9	10	39	78
logaritmus	7	8	6	9	10	40	80
humanismus	10	7	10	10	10	47	94
úsp(*m_typ*)	35	53	39	49	60		
úsp(*m_typ*) [%]	58	88	65	82	100		

Úspěšnost rozpoznání typu mluvčího se u jednotlivých subjektů mužského pohlaví (tabulka 10) pohybuje od 75 procent (mluvčí *m*5, základní tón F_0 se pohybuje v rozmezí 130–170 Hz, v patnácti případech z šedesáti klasifikován jako *žena*) do 100 procent (mluvčí *m*1 a *m*4, základní tón F_0 se u obou pohybuje v rozmezí asi 110–130 Hz). Všechny chybné výsledky v případě osob mužského pohlaví jsou způsobeny překročením ostré hranice intervalu rozhodujícího pravidla pro určení typu mluvčího. U slova „sinus“ (tabulka 10) lze pozorovat nejnižší celkovou úspěšnost (86 %), to je zřejmě způsobeno menším zastoupením znělé části ve slově, oproti slovům

s převažující znělostí (např. „humanismus“, typ mluvčího rozpoznán v 98 % počtu měření). V případě osob ženského pohlaví (tabulka 11) se úspěšnost rozpoznání typu mluvčího u jednotlivých subjektů pohybuje od 58 procent (mluvčí *m*6, základní tón F_0 kolísá v rozmezí 210 Hz až 270 Hz, v závislosti na intonaci promluvy. Při překročení hodnoty 255 Hz je osoba vyhodnocena jako *dítě*, a to ve 25 z celkových 60 slov) do 100 procent (mluvčí *m*10). Ve většině případů jsou u mluvčích ženského pohlaví chybné výsledky zapříčiněny překročením horní hranice intervalu F_0 (mluvčí klasifikován jako *dítě*). Pouze u osoby *m*7 došlo v několika případech k chybným výsledkům rozpoznání na základě záměrného poklesu intonace, tedy F_0, na hodnotu méně než 165 Hz (mluvčí klasifikována jako *muž*). Stejně jako u osob mužského pohlaví, tak i v tomto případě bylo dosaženo nejmenší úspěšnosti rozpoznání typu mluvčího prostřednictvím slova „sinus“ (úspěšnost 62 %), největší pak u slova „humanismus“ (úspěšnost 94 %).

8 Závěr

V teoretické části diplomové práce jsou popsány mechanismy tvorby řeči a základní metody úpravy řečového signálu za účelem dalšího zpracování v čase i frekvenci. Dále jsou popsány základní typy detektorů řeči pracující v časové a kepstrální oblasti. Velmi stručně je zmíněna existence detektorů založených na tzv. skrytých Markovovských modelech (*HMM*). Dále pak jsou popsány základní metody klasifikace v oblasti rozpoznání řeči, a sice vektorová kvantizace, *k-NN* klasifikace a algoritmus *DTW*. V poslední kapitole teoretické části jsou popsány metody určení základní frekvence hlasivek F_0, prostřednictvím autokorelační funkce a kepstrální analýzy. Teoretická část je zakončena stručným popisem mezinárodních fonetických abeced *IPA* a *SAMPA*.
V sedmé kapitole zabývající se popisem praktického návrhu systému pro detekci řeči jsou podrobně popsány jednotlivé funkce programu a používané výpočetní metody. Implementované funkce jsou popsány formou části jejich kódu nebo pomocí vývojových diagramů. V prostředí programu MATLAB byl vytvořen interaktivní systém pro zpracování řečového signálu. Pro detekci řeči ve zvukové nahrávce byl zvolen intenzitní detektor, schopný detekovat řeč v prostředí s přítomností hluku stacionárního charakteru. Minimální celkový poměr *SNR* pro záruku kvalitní detekce byl experimentálně vyhodnocen na 5 dB. Pod touto hranicí již detektor není schopen některé části promluvy odlišit od hluku.
Úspěšnost odhadu typu mluvčího dle kepstrální metody byla vyhodnocena na základě opakovaných měření u deseti jedinců (5 žen a 5 mužů). Minimální hodnota úspěšnosti odhadu typu mluvčího je u jedinců mužského pohlaví 75 %, u žen pak 58 %. Výsledky jsou u každého z mluvčích ovlivněny zejména variabilitou F_0, z jejíž průměrné hodnoty je určen daný typ mluvčího. Nejúspěšnější celkové výsledky odhadu typu mluvčího byly v případě mužů i žen získány řečením slova „humanismus“, kde dominuje znělá část promluvy (úspěšnost u mužů 98 %, u žen 94 %). Nejméně úspěšný odhad typu mluvčího náleží slovu „sinus“ (u mužů 86 %, u žen 62 %), jehož znělá část je oproti slovu „humanismus“ výrazně kratší.
V případě rozpoznávání izolovaných slov metodou *DTW* se na míře úspěšnosti rovněž podílí individuální barva hlasu, intonace, počet a podobnost slov ve slovníku vzorových promluv(slova *ahoj*, *olej*, *sinus*, *parametr*, *logaritmus*, *humanismus*). Minimální úspěšnost rozpoznání slova je u mužské části reprezentována 42 procenty u slova *ahoj* a maximální úspěšnost 96 procenty u slova *humanismus*. U žen je slovo *ahoj* rozpoznáno pouze ve 14 procentech případů měření, zatímco slovo logaritmus v 90 procentech počtu měření.
Celková úspěšnost odhadu konkrétní samohlásky na základě hodnot formantových frekvencí F_1, F_2 a počtu průchodů nulou se pohybuje v rozmezí od 66 % (u samohlásky „a“) do 98 % (u samohlásky „i“), přičemž úspěšnost odhadu všech hlásek konkrétní osobou se pohybuje od 64 % do 90 %.

9 Seznam použité literatury

[1] PSUTKA, Josef. Komunikace s počítačem mluvenou řečí. Praha: Academia, 1995. 287 s. ISBN 80-200-0203-0.

[2] MCLOUGHLIN, Ian. Applied Speech and Audio Processing. Leiden: Cambridge University Press, 2009. 216 s. ISBN 978-0-521-51954-0.

[3] VONDRA, Martin. Kepstrální analýza řečového signálu. In: Vysoké učení technické v Brně [online]. 2001 [cit. 2013-12-20]. Dostupné z: http://www.elektrorevue.cz/clanky/01048/index.html

[4] ČERNOCKÝ, Jan. Zpracování řečových signálů – studijní opora. In: Vysoké učení technické v Brně [online]. 2006 [cit. 2014-01-01]. Dostupné z: http://www.fit.vutbr.cz/study/courses/ZRE/public/opora/zre_opora.pdf

[5] PORUBA, Jiří, MATĚJÍČEK, Lukáš. Odfiltrování rušivých signálů ze zašumělé řeči. In: Vysoké učení technické v Brně [online]. 2002 [cit. 2013-12-11]. Dostupné z: http://www.elektrorevue.cz/clanky/02047/index.html#Kap2.2.1

[6] BEROUTI, M., R. SCHWARTZ a J. MAKHOUL. 1979. Enhancement of speech corrupted by acoustic noise. *ICASSP '79. IEEE International Conference on Acoustics, Speech, and Signal Processing* [online]. Institute of Electrical and Electronics Engineers, : 208-211 [cit. 2014-09-11]. DOI: 10.1109/ICASSP.1979.1170788. Dostupné z: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1170788

[7] HUANG, Xuedong. 2001. *Spoken language processing: a guide to theory, algorithm, and system development* [online]. Vyd. 1. New Jersey: Prentice-Hall [cit. 2014-01-01]. ISBN 01-302-2616-5.

[8] MOHYLOVÁ, Jitka a Vladimír KRAJČA. 2006. *Zpracování biologických signálů* [online]. 1. Ediční středisko VŠB – TUO [cit. 2014-02-08]. ISBN 978-80-248-1491-9.
Dostupné z: http://www.elearn.vsb.cz/archivcd/FEI/ZBS/Mohylova_Zpracovani%20biosignalu.pdf

[9] ČERNOCKÝ, Jan. 2009. Rozpoznávání pomocí DTW a HMM. *Brno University of Technology: Faculty of Information Technology* [online]. Brno [cit. 2014-08-08]. Dostupné z: http://www.fit.vutbr.cz/study/courses/ZRE/public/labs/05_dtw_hmm/05_dtw_hmm.pdf

[10] IVANECKÝ, Jozef. 2008. *Automatická transkripcia a segmentácia reči* [online]. Košice [cit. 2014-10-12]. Dostupné z: http://www.ivanecky.sk/Publikacie/dizertacka.pdf. Dizertačná práca. Technická Univerzita v Košiciach.

[11] BIRKHOLZ, P. 2013. *Modeling Consonant-Vowel Coarticulation for Articulatory Speech Synthesis*. Germany: An Open Access Biomedical Image Search Engine. Dostupné také z: http://openi.nlm.nih.gov/detailedresult.php?img=3628899_pone.0060603.g001&req=4

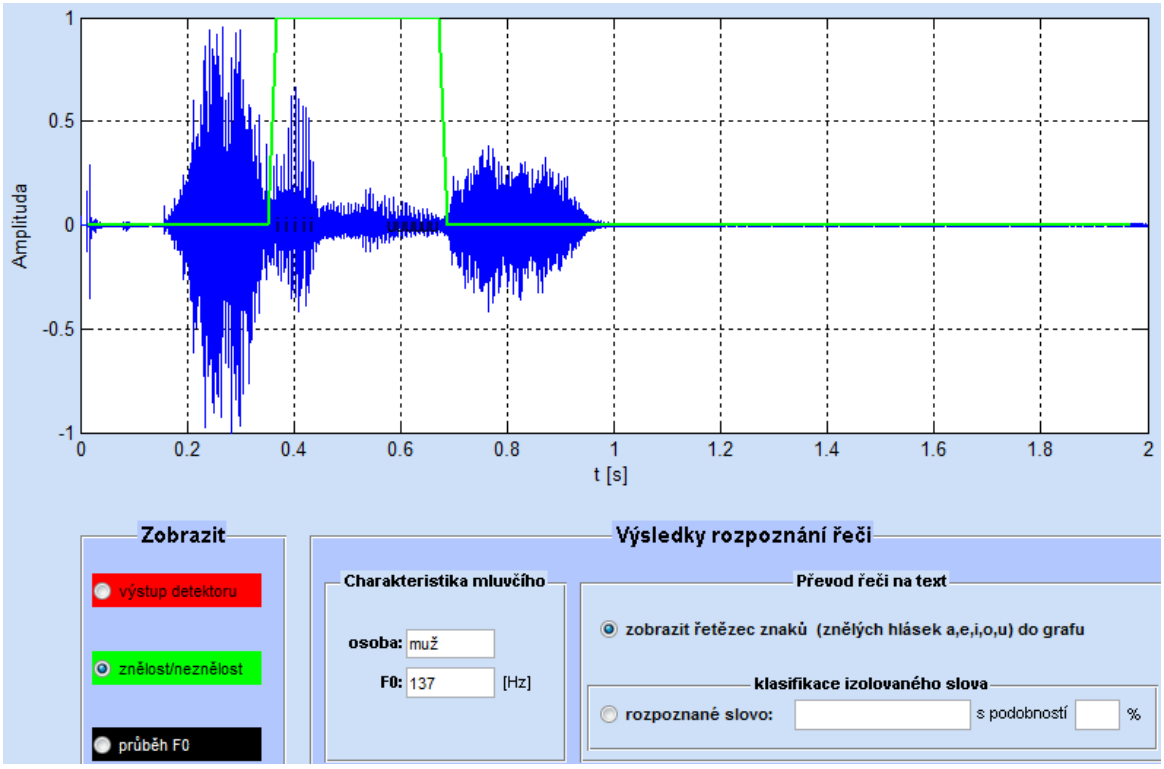
[12] The International Phonetic Alphabet (Revised to 1993). 2015. *Wikimedia Upload* [online]. [cit. 2015-04-03].
Dostupné z:
http://upload.wikimedia.org/wikipedia/commons/a/a5/IPA_Chart_Rev_1993.png

10 Seznam příloh

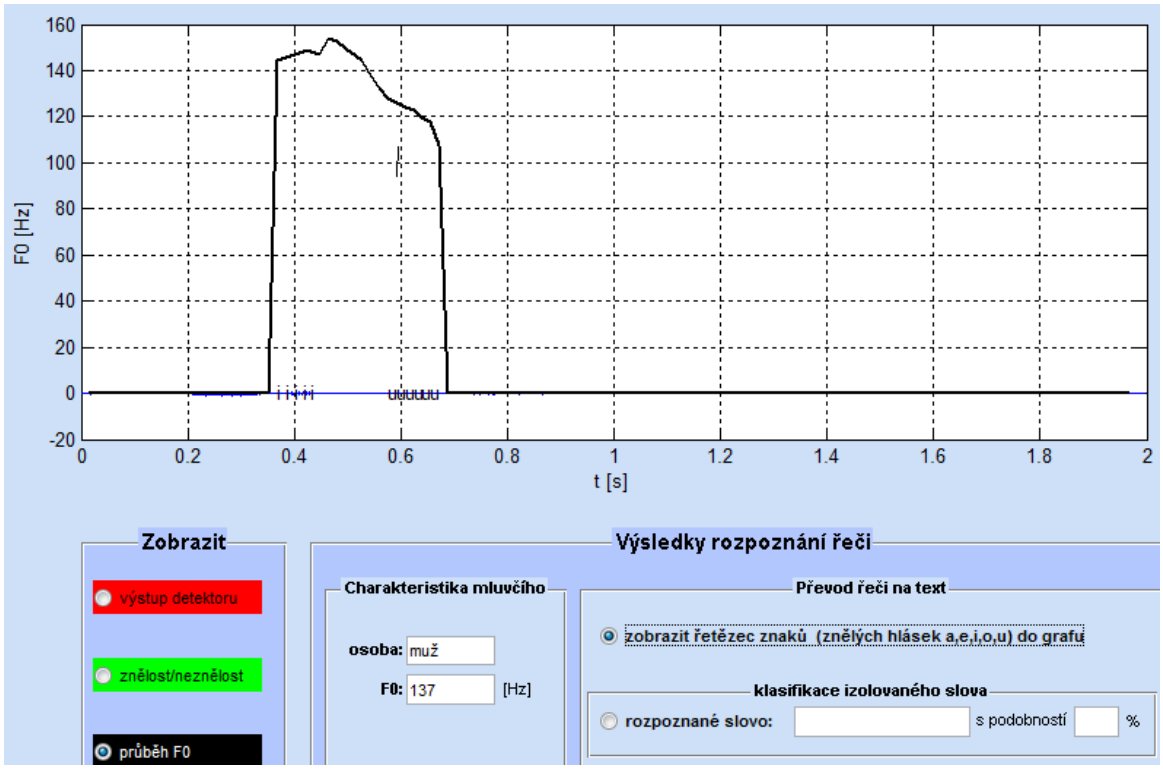
Příloha I Analýza znělosti a průběhu F_0 v čase.

Příloha II Detekce samohlásek s výpisem řetězce znaků.

Příloha I Analýza znělosti a průběhu F_0 v čase.



Zvýraznění znělého úseku slova „sinus“ (zeleně vyznačena část „inu“).



Časový průběh základní frekvence F_0 u slova „sinus“ (černě).

Příloha II Detekce samohlásek s výpisem řetězce znaků.